Universidad Autónoma de Madrid
Departamento de Biología Molecular

Structural analysis of macromolecular nanomachines by 3D electron microscopy: managing flexibility and heterogeneity in some defined cases

- TESIS DOCTORAL –

Ana Lucia Álvarez Cabrera

Madrid, 2015

Universidad Autónoma de Madrid

Departamento de Biología Molecular

Facultad de Ciencias

Universidad Autónoma de Madrid

Structural analysis of macromolecular nanomachines by 3D electron microscopy: managing flexibility and heterogeneity in some defined cases

Memoria presentada para optar al grado de Doctor en Bioquímica, Biología Molecular, Biomedicina y Biotecnología (Biociencias Moleculares)

Ana Lucia Álvarez Cabrera

DIRECTORES DE TESIS:

Dr. José María Carazo García
Centro Nacional de Biotecnología-CSIC

Dr. Carlos Oscar Sorzano Sánchez

Centro Nacional de Biotecnología -CSIC

CENTRO NACIONAL DE BIOTECNOLOGIA

CSIC

El trabajo recogido en esta memoria ha sido realizado en el Centro Nacional de Biotecnología (CNB-CSIC) bajo la dirección conjunta de los Drs. José María Carazo García y Carlos Oscar Sorzano Sánchez. Su financiación corrió a cargo de una beca predoctoral de Formación de Personal Investigador (FPI) y de los proyectos Instruct, CAM(S2010/BMD-2305), AIC-A-2011-0638 y AIC-A-2011-0638.

A las personas más importantes en mi vida: mis padres, mis hermanos y mi pareja.

Agradecimientos

Quisiera agradecer de forma especial a mis directores de tésis, José María Carazo y Carlos Oscar Sorzano, por la oportunidad de formar parte de este maravilloso grupo de investigación y, sobre todo, por su constante ayuda y apoyo a lo largo de estos años. También agradezco la compañía y enseñanzas de todos mis amigos y compañeros del CNB y de otros centros de investigación por los que he pasado durante este tiempo.

Table of contents

Figure index

Table index

Abbreviations

ΔZBD: Zinc binding domain deletion

2D: Two-dimensional

3D: Three-dimensional

aa: Amino acids

AAA+: ATPase associated with diverse cellular activities

adPEO: Autosomal dominant progressive external ophthalmoplegia

ATPγ-S: Adenosine-5'-(3-thiotriphosphate)

CENPJ: Centromere Protein J

CPAP: Centrosomal P4.1 associated protein

CL2D: Clustering two-dimensional classification

CTD: C-terminal domain

DNA: deoxyribonucleic acid

dTTP: Deoxythymidine triphosphate

dTDP: Thymidine diphosphate

dsDNA: Double-stranded DNA

EM: Electron microscopy

EMMDB: Electron Microscopy Data Bank

FSC: Fourier Shell Correlation

hp: hairpin

His-tag: Histidine tag

IMAC: Immobilized metal affinity chromatography

MCM: Minichromosome maintenance

MCM2-7: MCM2, MCM3, MCM4, MCM5, MCM6 and MCM7

MCM4/6/7-x2: MCM4/6/7 hexameric ring, formed by two MCM4/6/7 heterotrimers

MCPH: Primary autosomal recessive microcephaly

MT: Microtubule

MTOC: Microtubule organizing centre

mtSSB: Mitochondrial single-stranded DNA-binding

mtRNA: Mitochondrial ribonucleic acid

mtRNApol: Mitochondrial ribonucleic acid polymerase

MW: Molecular weight

n-1: All but one

NS: Negative staining

NS-EM: Negative staining electron microscopy

NTD: N-terminal domain

OCCM: ORC-Cdc6-Cdt1-(MCM2-7)

ORC: Origin recognition complex

PDB: Protein Data Bank

PCM: Pericentriolar material

pol γ: Polimerase gamma

pre-RC: Pre-replicative complex

RANSAC: Random Sample Consensus

RCT: Random Conical Tilt

RNA: Ribonucleic acid

RPD: Ribonucleic acid polymerase domain

SANS: Small-angle neutron scattering

SAXS: Small angle X-ray scattering

SEC: Size exclusion chromatography

SF4: Superfamily 4

SF6: Superfamily 6 (of helicases)

ssDNA: Single-stranded DNA

T7 gp4: Bacteriophage T7 gen 4

TEM: Transmission electron microscopy

WT: Wild type

ZBD: Zinc binding domain

Summary

Macromolecular complexes are truly nanomachines that play a crucial role in most cellular processes. Malfunctions of these biological machines are implicated in a wide variety of diseases, being the determination of their three-dimensional (3D) structures of high relevance in the medical and pharmaceutical fields. Indeed, an important part to understand the mechanism of a biological macromolecular assembly is the characterization of its structure and the interpretation of the potential dynamics that could represent the existence of multiple conformations.

In the present work there are presented the structural characterizations of three different biological cases: human CPAP, a centrosomal protein fundamental for centriole assembly; mouse MCM4/6/7, a group of proteins that form hetero-oligomeric ring-like complexes showing helicase activity; and human Twinkle, the mitochondrial replicative DNA helicase. All three cases share in common the fact of being highly dynamic and flexible proteins able to form different oligomeric complexes. Through each of the presented cases of study, it was addressed, to a varying extent, the challenge of elucidating the structures of a variety of oligomeric complexes, dealing with the conformational flexibility of these macromolecular machines at the same time that it was analyzed the possible biological/functional implications of the obtained results. For this goal, it was used a structural approach employing Transmission Electron Microscopy (EM) and computational analysis techniques. The architecture characteristics of the obtained two-dimensional (2D) images and three-dimensional (3D) models, along with the already reported functional, biochemical and structural information for the specific proteins, allow us to propose putative sequences of conformational/oligomeric changes. The presented density maps, together with the suggested associated structural transition pathways provide new insights into the possible mechanistic behavior of these machines of life.

Resumen

Los complejos macromoleculares son verdaderas máquinas nanoscópicas que juegan un papel crucial en la mayoría de los procesos celulares. El mal funcionamiento de estas máquinas biológicas está relacionado con un gran número de enfermedades, por lo que la determinación de sus estructuras tridimensionales (3D) resulta de gran importancia con fines aplicados a los campos de la medicina y la farmacéutica. Para llegar a comprender el mecanismo de los ensamblajes moleculares biológicos se requiere, en parte importante, de su caracterización estructural y de la dinámica que puede representar la existencia de múltiples conformaciones.

En este trabajo se presentan las caracterizaciones estructurales de tres casos biológicos distintos: la proteína centrosomal de humano CPAP, esencial para el ensamblaje de los centriolos; la helicasa MCM4/6/7 de ratón, que consiste en un grupo de proteínas que forman complejos en forma de anillo, capaces de separar DNA de doble cadena; y Twinkle, la helicasa de ADN mitocondrial de humano. Todos los tres casos comparten en común la característica de ser proteínas muy flexibles y altamente dinámicas, capaces de formar distintos tipos de complejos oligoméricos. A través de cada uno de los casos presentados, se abordó, en mayor o menor grado, el reto de elucidar las estructuras de una variedad de complejos oligoméricos, lidiando con la flexibilidad conformacional de estas máquinas macromoleculares al mismo tiempo que se propuso el posible significado biológico/funcional de los resultados obtenidos. Con este propósito, se empleó una aproximación estructural mediante estudios por técnicas de microscopía electrónica de transmisión (TEM) y análisis computacional. Las características estructurales de las imágenes bidimensionales (2D) y los modelos tridimensionales (3D), junto con la información funcional, bioquímica y estructural que ha sido reportada para cada proteína específica, permitió proponer secuencias putativas de cambios conformacionales/oligoméricos. Los mapas de densidad presentados y los respectivos pasos de transición estructural propuestos proporcionan nuevas perspectivas en el posible comportamiento mecánico de estas maquinas de la vida.

Chapter 1

1. Introduction

1.1. Thesis Outline

This thesis is divided into six main sections as follows:

1. In the introductory chapter, basic concepts about protein structure are briefly summarized as well as the most common experimental techniques used in the field of structural biology to elucidate the 3D architecture of macromolecules. A special emphasis is placed into single particle electron microscopy (EM), the technique used for the resolution of the different structures obtained during the development of the three subprojects that constitute this thesis. In order to contextualize the biological relevance of each of these subprojects, the processes into which each of the related proteins/complexes are known to be involved are explained as well as what is currently known about their own properties and functional roles.

2. The objectives of the thesis are presented in the second section.

3. The third chapter describes the materials and methods.

4. The fourth section is the presentation of results. This part is subdivided into three different but related sub-sections from the perspective of managing biological samples characterized by both a high degree of conformational flexibility and oligomeric heterogeneity. Here the novel structural information obtained in all the cases under study is shown.
   4.1. The first project is focused on the human centrosomal protein CPAP. It encompasses the molecular cloning, expression and purification of a N-terminal truncate construct of the protein, as well as the acquisition and processing of EM images of multiple oligomeric species formed by this protein. It is proposed a putative sequence of increasing oligomerization steps, as well as the possible visual identification of the largest supramolecular CPAP structures within the centriole.

4.2. The mouse hexameric helicase MCM4/6/7 is the subject of the second case of study. Despite being an enormous challenge from the image analysis point of view, the structural dynamism of this sample was a unique opportunity to capture conformational and oligomeric changes that could give new insight into the assembling and action mechanisms of this complex.

4.3. The last project involves the work with purified samples of TWINKLE, the human mitochondrial DNA helicase. In a first instance the conditioning and selection of the optimal sample conditions for its later use in EM studies was carried out. The 2D image analysis and the initial 3D models of different oligomeric and conformational states of the protein are presented.

5. In this part, the results of each of the projects are discussed, placing them into the context of the *state of the art* of its related topics, including an analysis of the relevance and possible contributions to the biological and biomedical fields.
6. The final section summarizes the main conclusions and findings of this thesis.

1.2. Structural biology

Structural biology is a branch of the biological sciences that incorporates the principles of molecular biology, biochemistry and biophysics, among other disciplines, including bioinformatics, to study the molecular structure and dynamics of biological macromolecular assemblies, in particular proteins and nucleic acids, and how alterations in their structures affect their function.

1.2.1. Protein structure

Proteins are large and complex molecules made up by a linear sequence of amino acids, whose different levels of organization in the three-dimensional (3D) space (Figure 1), give rise to structures involved in the execution and direction of virtually all the processes required to maintain living systems. Some proteins can form multimers with the same or different type of proteins (as well as other chemical entities, like nucleic acids or sugars), and such supramolecular assemblies constitute the final functional units.

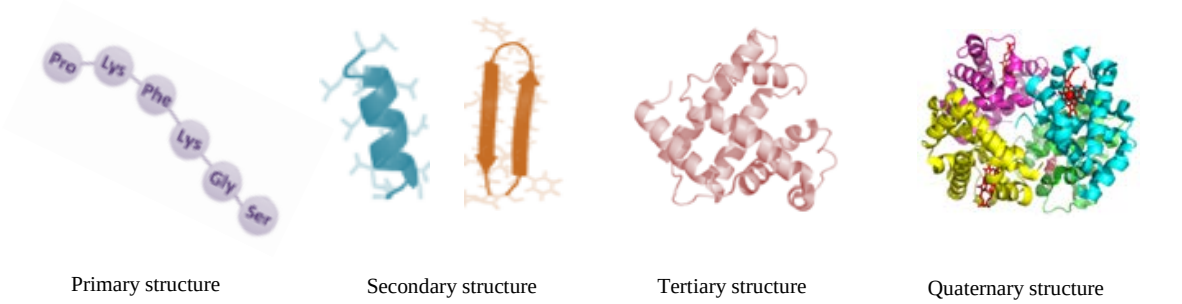


Figure 1. Structural levels of protein organization. Primary structure (amino acid sequence), secondary structure (in blue an α-helix and in orange a β-sheet), tertiary structure (the fold of a chain), and quaternary structure (multi-subunit complex made of several combined chains).

Proteins can be broadly classified according to their shape and solubility into three major classes:

- **Fibrous proteins** are poorly soluble, but highly flexible rod-like structures, where the secondary structure is the predominant feature and little or no tertiary structure is displayed.
- **Globular proteins** are more compact and water soluble molecules, since its structure is essentially organized in a way that most of its hydrophobic amino acids form an internal core that is sheltered by polar residues exposed to the watery surface.
- **Membrane proteins** are relatively flexible and aqueous insoluble molecules that are found in close association with lipid membranes through their hydrophobic surfaces, while their hydrophilic residues tend to protrude out.

Nevertheless, describing the structure of a protein might not be so straightforward, since a unique and static structure is not a realistic situation for most proteins. Besides, there are proteins whose sequences are predicted to be partially or completely unstructured (Romero & Obradovic 1998). Although intrinsically disordered proteins (IDPs) or regions (IDRs) can acquire a more stable folding upon interacting with a binding partner, it has been observed that some ID segments retain a disordered state in the new formed complex, which is known as a fuzzy complex (Tompa & Fuxreiter 2008).

1.2.2. Protein dynamics and structural heterogeneity

For decades ago it was assumed that each protein presented a unique functional 3D structure and that its interaction with a partner followed a highly specific and rigid lock-and-key model. However, far from being static, isolated objects, proteins are dynamic entities that make use of different levels of structural flexibility (i.e. local, regional or global) to interact or associate with other molecules in order to carry out their function. While it is true that activity of many proteins is associated with a stable 3D architecture, their function must still be accomplished by a dynamic sequence of movements and biochemical interactions occurring at different timescales and ranging from few atomic fluctuations to drastic conformational rearrangements, which can include its assemble into higher order macromolecular structures. It is not unusual to find proteins or complexes that exist simultaneously in a variety of functional conformations. The conformational and oligomeric dynamic behavior shown by some proteins brings a broad and heterogeneous set of structures, whose elucidation would constitute an important pillar for the understanding of their biological function.

1.2.3. Protein structure determination

A comprehensive study of protein structure and its dynamics is a determining factor for a successful and deeper understanding of their functions, so describing the conformational states of proteins and biological macromolecular assemblies is a critical point in the structural biology field. The architectural description of heterogeneous samples (due to protein flexibility and/or different oligomerization states) is a challenging, yet exciting, task for structural biologists, who often employ different experimental and bioinformatics tools to handle this kind of problems.

1.2.3.1. Experimental methods

To date, the most used and reliable experimental approaches for structural determination of macromolecules are X-ray crystallography, nuclear magnetic resonance (NMR) spectroscopy, which is mainly employed for high resolution model of small proteins, and single particle electron microscopy (EM) analysis to obtain medium-resolution density maps of larger proteins and complexes. Each of these methods has its own advantages and limitations, so an integrated strategy using information obtained from different techniques, when available, must be followed

in order to reach the atomic details for any biologically significant state of a macromolecular assembly.

X-ray crystallography

This crystallographic technique employs X-rays with a wavelength around 1Å, which allows resolving structures at the atomic level. Obtaining a good protein crystal is the most crucial and difficult task of the whole process, so a number of hurdles must be overcome before getting a X-ray diffraction pattern from which the structural information of the molecule could be extracted. A first bottleneck is to have a considerable amount of a highly pure, homogenous and concentrate sample, which may be very complicated in a large number of cases. A second, and no less complicated step, is the production of a large enough (a minimum of 0.1-0.2 mm in at least 2 dimensions) and nicely organized crystal diffracting at good resolution (Egli 2010). Crystal nucleation and growth are non-trivial challenges, since those are extremely complex and poorly understood processes. Proteins presenting IDRs or very flexible domains are frequently incompatible with the X-ray crystallography technique. On the other hand, proteins forming part of a crystal have been forced to acquire the organized arrangement that allows its crystallization (Figure 2), so the resolved structure may not necessary reflect a physiological conformation of the particle.

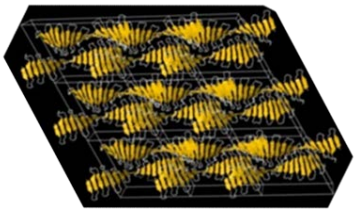


Figure 2. Structure of the wild-type TCP10 domain of *Danio rerio* CPAP (PDB_4BXP) as obtained from a crystalline sample by X-ray crystallography.

Nuclear magnetic resonance (NMR) spectroscopy

The methodological implementation of NMR involves the use of a number of separate and specialized techniques, and is traditionally used for the determination of the atomic structure and dynamic of relatively small proteins or nucleic acids in solution. As in large molecules the NMR signals are more prone to overlap and their lower frequency Brownian motion produces a

significant reduction in sensitivity and resolution, typically the samples employed do not exceed 30-40kDa. Through NMR spectroscopy studies, a set of structures coming from an analyzed macromolecule can be obtained (Figure 3), taking advantage of the fact that the nuclei of atoms can be aligned by applying a strong magnetic field, and then the nuclei magnetic momentum can be changed back and forth from the equilibrium by using the energy of certain radio wave frequencies. The detected nuclear resonance spectrum is interpreted to obtain structural information.

Figure 3. NMR structure (PDB_2M45) of the C-terminus of the minichromosome maintenance protein MCM from Sulfolobus solfataricus.

Although in recent years the use of more powerful technologies and new complementary methods has allowed the use of NMR spectroscopy in the study of some larger proteins and complexes (Frueh et al. 2013), those are still exceptional cases. As in X-ray crystallography, a crucial requirement for NMR experiments is the high purity of the protein, so this technique is limited by both, size and sample availability.

Transmission Electron Microscopy (TEM)

EM is a versatile technique that can be employed in the structural study of a broad variety of samples types and sizes. For example, *single-particle EM* is the method of choice when working with individual particles and filaments with or without symmetry, which have not been able to be solved by other techniques. *Electron crystallography* is employed for the study of small 2D crystals (< 0.1 um) because they cannot be analyzed by X-ray crystallography due to their limiting size. Finally, *electron tomography* allows obtaining 3D information of pleomorphic entities, such as cells or of sub-cellular macromolecular objects.

In an electron microscope, electrons provided by an emission source (e.g. a tungsten, lanthanum hexaboride filament, or a field emission gun -FEG-) are accelerated along a vacuum system column

by an electric field at a high voltage and are focused by electromagnetic lenses into a collimated beam. The electron beam passes through the sample interacting with the specimen, and the transmitted electrons that will finally reach a detector form a 2D image of the sample being imaged. The brightness of the different areas in the image is proportional to the phase shift between scattered and unscattered electrons transmitted through the sample at each point. Under the electron microscope, biological samples produce a very low contrast and are highly sensible to damage caused by vacuum conditions and electron dose irradiation, so the sample structure must be preserved in some way before and during its observation. Two main techniques for sample preparations can be employed, although variations of them can be done:

- Negative staining (NS) consists in the incubation of the sample with an electron dense material (usually a solution of a heavy metal salt), which covers the particles creating a sort of mask with the same shape, that is more resistant to radiation and dehydration. The stain increases the Signal-to-Noise Ratio (SNR) of the images, but as a drawback, the resolution is limited by the grain size and penetration capability of the stain; additionally, some fragile structures can be deformed. Particles as small as 53 kDa (Zhang et al. 2013) have been solved by NS, although currently, this small size is more the exception that the norm.
- Cryo-vitrification of the sample is carried out by its fast immersion in liquid ethane; in this way the specimen gets frozen into an amorphous layer of ice that keeps the particles in a state close to the physiological one. Vitrified samples do not suffer from the resolution limit of NS, but the contrast is very low, which is a fundamental issue when further processing these cryo-EM images.

Detectors for image acquisition were traditionally photographic films or charge-coupled devices (CCD), but recently, a new generation of Direct Electron Detectors (DED) has been developed which directly receive the energy of the incident electrons. Among other advantages, they greatly improve the spatial and time resolution of the image (Milazzo et al. 2011). The posterior 3D reconstruction method can be done thanks to the fact that both phase and amplitude information is present in the 2D image. Resolution of EM density maps is normally measured by random but equally dividing the image data set into two sub-sets, so two independent 3D reconstructions are obtained and a comparison between their Fourier shell correlations (FSC) can be carried out. The most commonly used criterion is the 0.5 FSC cutoff, which refers to when the correlation

coefficient of the Fourier shells drops below 0.5 (Manuscript 2011), although different criteria are also possible (Rosenthal & Henderson 2003), (Scheres & Chen 2012).

Single-particle EM

Most of the activities within the cell are carried out by dynamic interactions and association between proteins. The hallmark strength of TEM, compared to X-ray crystallography and NMR, is the capacity of this technique to capture structural information of large biological macromolecules and complexes in their native environment, using just a tiny amount (nanograms) of sample. Sample purity requirements do not need to be so high, especially if the particle under study can be visually distinguished from possible contaminants. Single-particle EM analysis is a powerful tool for the structural study of macromolecular assemblies that exist in multiple conformational states, which confer to this technique an additional and unique advantage when compared with other methods. Big size, symmetry and homogeneity of the analyzed particle, are factors that boost the reachable resolution of the final 3D reconstruction, while the opposite situation makes more difficult but does not preclude obtaining volumes with valuable structural features (Burgess et al. 2004).

In single-particle analysis, thousands of preprocessed images obtained from randomly oriented particles, are computationally aligned and classified. Those images are subsequently used for constructing initial density maps (using methods like random conical tilt -RCT- (Radermacher 1988) or RANSAC (Vargas et al. 2014)) or for the refinement of preexistent volumes (Frank 2006). The above mentioned processes use to be carried out by employing a combination of different software image processing programs (e.i. Xmipp (De la Rosa-Trevín et al. 2013), Relion (Scheres 2012), EMAN (Ludtke et al. 1999), FREALIGN (Grigorieff 2007), among others).

1.2.3.2. Complementary bioinformatics methods

Currently, a number of technical and sample related issues are still limiting factors for EM to reach the theoretical resolution that electron scattering would allow, especially for small, asymmetric particles (Glaeser & Hall 2011). Resolution can be partially improved by employing hybrid methods, that is, a process that involves fitting in the EM density map, the high resolution

structures solved by other techniques (Rossmann et al. 2005). The level of success of this method depends on the actual resolution of the EM map and the coverage of the available atomic models to be fitted. Bioinformatics programs used for *de novo* predicting and modeling atomic structures are very helpful when there are not experimentally determine structures, or when the available ones are incomplete. Prediction of secondary structures, globular or disordered regions based on the protein sequence, also brings useful information.

1.3. Biological system: CPAP centrosomal protein

1.3.1. The centrosome

Centrosomes are found in most animal cells as the primary microtubule organizing centers (MTOC), carrying out an important number of cellular processes, such as the regulation of normal cell division and motility; the organization of interphasic cytoskeleton's microtubules and the spindle apparatus, and the establishment of the cellular polarity. This organelle is composed by two barrel-shape centrioles (a mature one called the mother centriole, and an immature centriole that is assembled during the previous cell cycle, called the daughter centriole) interconnected through fibers attached to their proximal base, and surrounded by the pericentriolar material (PCM), an apparently amorphous proteinaceous matrix (Bettencourt-Dias & Glover 2007) (Figure 4A).

The centrosome is normally localized at the central region of interphasic cells. However, after its replication through the S phase of the cell cycle, the centrosomes migrate to opposite poles of the cell during the prophase of mitosis, allowing the formation of the mitotic spindle between the two centrosomes (Figure 4B). The cellular bipolarity and the equal partition of chromosomes into two daughter cells are all dependent on normal centrosome function and duplication, which at the same time depends on the correct assemble of the centrioles. Centrosomes are sites of integration and activation of proteins that trigger cell division. During the last several years numerous proteins predicted to be notably rich in coiled-coil motifs and unstructured regions have been reported to be associated with the centrosome, either as permanent components, or temporally associated and concentrated at the centrosome during some stages of the cell cycle (Nogales-Cadenas et al. 2009; Alves-Cruzeiro et al. 2014).

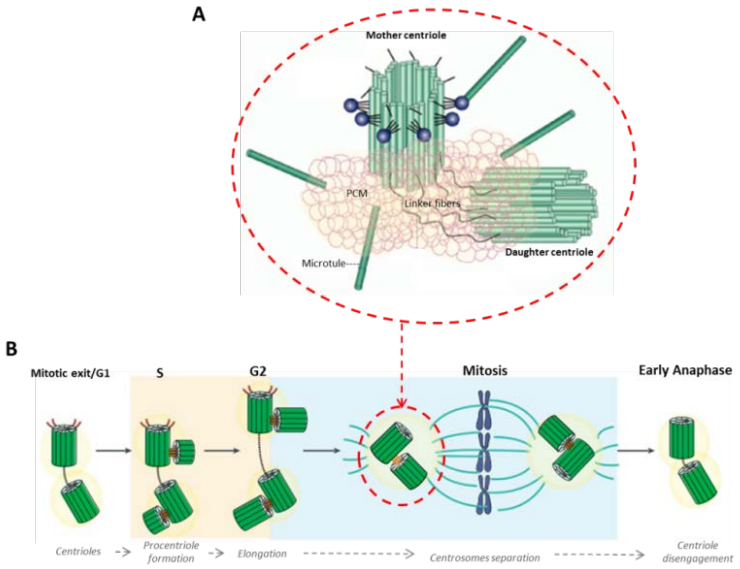


Figure 4. Centrosome. **(A)** Main structural components of centrosome. Adapted from (Doxsey 2001). **(B)** Representation of the centrioles duplication process coupled with the cell cycle. Adapted from (Debec & Sullivan 2010).

1.3.2. CPAP protein

CPAP (Centrosomal P4.1 associated protein, also known as Centromere Protein J –CENPJ-, or as Sas-4 in Drosophila and SAS-4 in *C. elegans*) is a centrosomal protein that plays a crucial role in the recruitment of microtubules (MT) during centriole formation, so it is a critical component implicated in the biogenesis of the centrosome (Cho et al. 2006),(Kohlmaier et al. 2009). It has been demonstrated that CPAP and its orthologs are mainly associated with the tethering of the PCM to centrioles (X. Zheng et al. 2014), (Kohlmaier et al. 2009; Tang et al. 2009; Tang et al. 2011), (Gopalakrishnan et al. 2011), (Hsu et al. 2008), being this a key processes for the proper assembly of the centrosome which, in turn, is tightly implicated in the regulation of cell cycle progression (Doxsey et al. 2005).

CPAP is regulated during the cell cycle (Tang et al. 2009) and can be localized in different structures within the centrosome (Kleylein-sohn et al. 2007). CPAP associates with many other centriolar and PCM proteins, such as P4.1R-135 (Hung et al. 2000), CEP152 (Cizmecioglu et al. 2010), CEP135 (Y.-C. Lin et al. 2013), (Vulprecht et al. 2012), CEP120 (Y.-N. Lin et al. 2013),(Comartin et al. 2013), Tubuline heterodimers (Hung et al. 2004), γ-tubulin (Hung et al. 2000), (Hung et al. 2004), STIL (Tang et al. 2011), (Cottee et al. 2013), 14-3-3 (Chen et al. 2006) and , Centrobin (Gudi et al. 2014). Furthermore, it has been reported that SAS-4 also interacts with

cytoplasmic complexes, and very interestingly, that its role in the PCM assembly (through its N-terminus) is independent to the role that it carries out along the formation of the centriole (through its C-terminus) (Gopalakrishnan et al. 2011). All the mentioned facts highlight the multifunctionality nature of CPAP, which make us realize about the potential structural complexity that this protein must have in order to be properly adapted to perform different but specific tasks in each situation when it is required. At present, the structural performance of CPAP continues poorly understood.

In CPAP, the non-conservative mutation E1235V has been described in patients diagnosed with primary microcephaly (MCPH) (Leal et al. 2003), a genetically heterogeneous disorder characterized by a significative reduction in normal brain volume and intellectual disability, whereas maintaining the general architecture of a normal brain. To date, fourteen loci, including the gene CENPJ that codes for CPAP (locus MCPH6), have been found to be related with MCPH disease (Table 1); most of these genes code for proteins located in the centrosome, either in a constitutive way or transitorily associated to the centrosome during part of the cell cycle.

Locus name	Protein
MCPH1	*Microcephalin 1
MCPH2	*WDR62
MCPH3	*CDK5RAP2
MCPH4	CASC5
MCPH5	*ASPM
MCPH6	*CPAP/CENPJ
MCPH7	*,ᴿ STIL
MCPH8	*,ᴿ CEP135
MCPH9/SCKL5	*,ᴿ *CEP152*
MCPH10	ZNF335
MCPH11	*PHC1*
MCPH12	*CDK6
---	*SAS-6
SCKL6	**CEP63*

* Proteins localized at the centrosome in a constitutive way or transitory associated to centrosome during some stage of the cell cycle. ᴿProteins that have been observed to directly interact with CPAP.

Table 1. Primary autosomal recessive microcephaly (MCPH) associated proteins.

1.3.2.1. CPAP structure

At the structural level, CPAP presents a C-terminus with several short glycine-rich repeats that together form the so called G-Box domain (also known as TCP domain), which has a highly conserved sequence through evolution, pointing to its primordial role in centriole/centrosome development, a fact that has been widely corroborated experimentally. In recent years, a number of crystallographic structures of the G-Box/TCP domain of CPAP have been resolved, revealing a solvent-exposed β-sheet structure that remains stable by its own despite lacking a hydrophobic core (X. Zheng et al. 2014; Hatzopoulos et al. 2013; Cottee et al. 2013). The rest of CPAP is formed by disordered regions and coiled-coil motifs (Hung et al. 2000) (Carvalho-Santos et al. 2010). The human CPAP (*Hs* CPAP) protein presents five coiled-coil domains, of which the 4^{th} and 5^{th} coiled-coils (CC4 and CC5, respectively) show a circular dichroism (CD) spectra consistent with an α-helical secondary structure (Hatzopoulos et al. 2013), which agree with the *in silico* prediction of its amino acid (aa) sequence. CPAP also contains both a MT binding domain (MBD, 423-607aa), that binds to a polymerized microtubule, and a MT destabilizing domain (MDD, 311-422aa), that binds to an α/β-tubulin heterodimer (Figure 5). The later interaction results in inhibiting nucleation and MT depolymerization (Hung et al. 2004), (Cormier et al. 2009), (Hsu et al. 2008).

The region of *Hs* CPAP between residues 897-1338 contains the G-Box (residues 1150-1338), the coiled-coils CC4 and CC5 (which together cover the residues 897-1056, and that we will call "CC5/CC4" from now on), and a predicted unstructured zone (residues 1066-1149, that will be referred from now on as the "CCGb-linker" region), that makes a linker between CC4/CC5 and G-Box (Figure 5A*). Interestingly, it has been found that a number of regions included in this 897-1338 sequence are involved in the interaction of CPAP with other centrosomal proteins, such as P4.1R-135, CEP135, CEP120, STIL and 14-3-3 (Figure 6). The CC5 domain is responsible for the dimerization of CPAP (Zhao et al. 2010). Indeed, it is known that the formation of CPAP oligomers as well as its disruption once the protein is phosphorylated during mitosis, is essential for the correct centrosome integration along the cell cycle (Zhao et al. 2010) and for the regulation of the interactions of CPAP with other proteins (Chen et al. 2006).

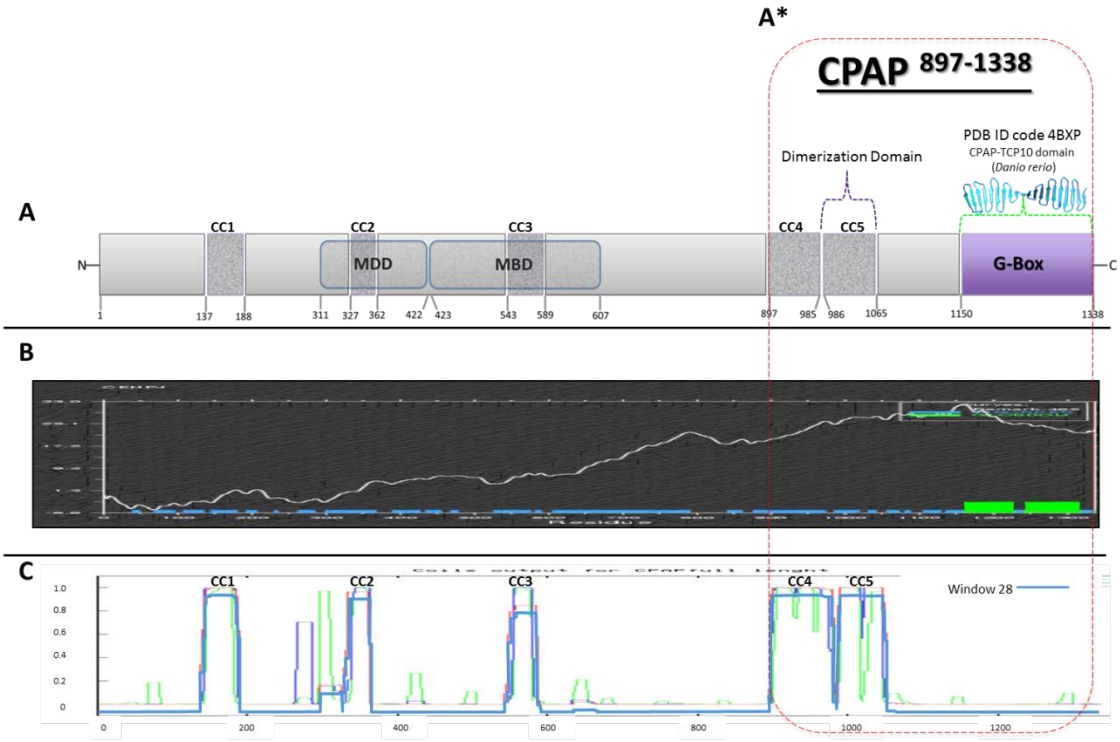


Figure 5. Structural domains in human CPAP protein. (A) Schematic representation of the structural domains in *Hs* CPAP full-lenght protein. Dark gray boxes represent the five predicted coiled-coils (labelled as CC1, CC2, CC3, CC4 and CC5), being the CC5 coiled-coil the oligomerization domain of CPAP; light gray corresponds to unstructured regions and the purple box refers to the G-Box domain, whose cristallographyc structure was already solved in *Danio rerio* (PDB-4BXP); MDD: Microtubule destabilization domain and MBD: Microtubule binding domain. **(A*)** Red dashed lines demark the CPAP$^{897\text{-}1338}$ construct. **(B)** In silico prediction of globular and disordered regions of proteins *Hs* CPAP (full-length), obtained using the program GlobProt2. The green boxes show the predicted globular domains, whereas the blue boxes correspond to disordered regions of the protein**. (C)** Probability prediction of five coiled-coil structures in CPAP, calculated with the software COILS.

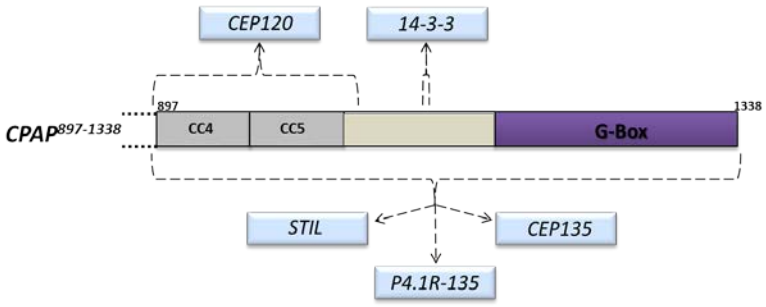


Figure 6. CPAP897-1338 interacting partners.Approximate regions within CPAP$^{897\text{-}1338}$ (demarked by the corresponding slashed curly brackets) that have been reported to interact with other centrosomal proteins (light-blue boxes).

Taking into account the current available functional experimental data of CPAP, its multiple interaction partners, and the peculiar supramolecular β-sheet structure of Its G-Box domain (Kirkham et al. 2003; Gopalakrishnan et al. 2011), It has been proposed that CPAP works initially as a scaffold for different cytoplasmic proteins, followed by tethering of the new ensemble PCM complexes to the centriole through its C-terminal domain, allowing to continue the formation of normal and functional centrosomes (Gopalakrishnan et al. 2011; Hatzopoulos et al. 2013; X. Zheng et al. 2014), (Tang et al. 2009). Despite of the biochemical and biophysical evidences of CPAP dimer formation (Hatzopoulos et al. 2013), (Zhao et al. 2010), there are neither clear structural data of such complex or reports of other organized higher order oligomeric arrangements of CPAP. Based on our work on the purification and EM image analysis of the CPAP897-1338, here we report, for the first time, that this protein acquires several oligomeric states, which reveals new and important structural features that could significantly contribute to the understanding of the biological role of CPAP.

1.4. Biological system: MCM4/6/7 helicase

1.4.1. MCMs proteins

Minichromosome maintenance (MCM) genes were identified for the first time in cold-sensitive mutant yeasts that showed abnormal plasmid segregation (defective minichromosome maintenance) (Maine et al. 1984), (Hennessy et al. 1991), (Moir & Botstein 1982). Early studies placed MCM proteins as central pieces in the initiation of the DNA replication process, a role substantially confirmed by subsequent work showing the helicase activity of MCMs ring complexes in different systems (for a review see (Bochman & Schwacha 2009), (Bell & Botchan 2013)). Despite being well known for their role as replicative helicases, other observations revealed that the number of copies of MCM proteins in the cell not only exceeds by far the number of replication origins, but also that they are widely distributed on heterochromatin and that their levels seems to remain stable even after the end of DNA replication (Adachi et al. 1997), (Dimitrova & Todorov 1999), (Krude & Musahl 1996), (Madine et al. 1995), (Kinoshita & Johnson 2004), (Claycomb et al. 2002), (Dalton & Whitbread 1995), (Forsburg et al. 1997), (Kimura et al. 1995), (Schulte et al. 1995), (Tsuruga et al. 1997). This gives rise to the so called "MCM paradox" (Woodward et al. 2006), (Ibarra et al. 2008), (Hyrien et al. 2003), (Takahashi et al. 2005), (Das et al. 2014), and opened the discussion on whether MCMs could be involved in other cell processes besides DNA replication. Indeed, to date, a number of studies suggest a participation of MCMs proteins in transcription and remodeling of chromatin (Yankulov et al. 1999), (DaFonseca et al. 2001), (Dziak et al. 2003), (Fitch et al. 2003), (Sterner et al. 1998), (Holthoff 1998), (Ishimi et al. 2001). It is also known that under conditions of cellular stress, the activation of dormant origins of replication depends on this "extra" amount of MCMs (Crevel et al. 2007), (Woodward et al. 2006), (Ibarra et al. 2008).

In the clinical field, MCMs have begun to be recognized by their potential as cell proliferation markers. Furthermore, their expression levels seems to be associated with several pathologies, which make them useful as a diagnostic and prognostic tool for different malignancies, such as cancer (Giaginis & Vgenopoulou 2010).

It seems evident that MCM proteins contribute in different ways to maintain the integrity of the eukaryotic genome and, consequently, to cell biogenesis, but further research is needed to understand other roles and processes where MCMs could be involved.

1.4.2. MCMs as helicases

Helicases have been classified into six superfamilies according to their sequence motif conservation. MCM proteins are found in both archaea and eukaryotes and they belong to the functionally heterogeneous AAA+ (ATPase associated with diverse cellular activities) superfamily 6 (SF6) of helicases. All SF6 member share similar motifs at their AAA+ "core domains" (Singleton et al. 2007), (Daniel et al. 2013). More specifically, MCM proteins are members of clade 7, that is characterized by the insertion of an additional α-helix (helix 2 insert) when compared to the general structurally conserved ATP-binding module of the AAA+ group (Iyer et al. 2004), (Erzberger & Berger 2006).

The MCM canonical architecture can be summarized by dividing the sequence into three general segments (Figure 7A):

- A N-terminal of low sequence preservation among MCM´s homologs, that binds both ssDNA and dsDNA, and strongly affects the helicase oligomerization and processivity (Kasiviswanathan et al. 2004), (Jenkinson & Chong 2006), (Bochman & Schwacha 2007), (Bae et al. 2009), (Pucci et al. 2007), (Liu et al. 2008). This región, in turn, is divided into three structural subdomains (sA, sB and sC)(Costa & Onesti 2009).
- A highly conserved AAA^+ core that is responsible for the catalytic ATPase and helicase activities (Tye & Sawyer 2000), (Brewster et al. 2008), (Bochman & Schwacha 2009). This domain has an α/β subdomain connected via the α/β- α linker to an α domain (Costa & Onesti 2009)
- A C-terminal domain that, despite of a low level of sequence conservation among MCM´s homologs, shares a pattern compatible with a helix-turn-helix (HTH) fold, a typical DNA-binding motif. In support of this observation, the atomic structures of the C-teminal domains of human (*Homo sapiens*, *Hs*) MCM6 (PDB-2KLQ), *Sso*MCM (PDB-22M45) and *Mth*MCM (PDB-2MA3), all present a HTH fold. The high flexibility of this part is likely to be the principal factor that makes its structural characterization difficult (Costa et al. 2006b), (Brewster et al. 2008), (Krueger et al. 2014). This domain shows a negative regulatory

effect on the helicase activity of MCM archaeal homologs (Aravind & Koonin 1999), (Jenkinson & Chong 2006), (Pucci et al. 2007), (Wei et al. 2010), (Barry et al. 2007).

Some eukaryotic MCM proteins display additional low conserved sequence extensions at their N- and/or C-terminus that may play specialized roles in particular organisms (Lyubimov et al. 2012), (Wei et al. 2010), (Xu et al. 2013), (Tye & Sawyer 2000), (Sun et al. 2014), (Sheu et al. 2014). In *Methanothermobacter thermautotrophicus* and *Sulfolobus solfataricus* (the most commonly studied archaea systems), there is only one class of MCM protein, *Mth*MCM and *Sso*MCM, respectively (Kelman et al. 1999), (McGeoch et al. 2005). In contrast with archaea, eukaryotes present several MCM homologous proteins, where each of the MCM2, MCM3, MCM4, MCM5, MCM6 and MCM7 genes defines a preserved separate family sharing significant sequence similarities to each other, mostly but not confined to the region that conforms the AAA^+ core domain (Koonin 1993).

Thanks to the high sequence conservation of the AAA^+ core domain across species and between MCM subfamilies, the analysis of the near full-length (lacking the first 6 N-terminal and the last 85 C-terminal residues) crystallographic structure of *Sso*MCM (PDB ID code 3F9V) has provided more detailed information about several characteristic and relevant elements of MCM helicases (Brewster et al. 2008) (Figure 7B). The most prominent features observed are two globular regions, a big AAA^+ domain and a smaller N-terminal domain, which are interconnected by a narrow linker, giving the appearance of a slim “waist” flanked by two asymmetric lobules. This broad shape can be distinguished in the side views of the 2D images and 3D maps of different MCM helicase complexes obtained to date by Transmission Electron Microscopy (EM) (Lyubimov et al. 2012), (Costa et al. 2011). The C-terminal domain is a much smaller and flexible part compared with both AAA^+ and N-terminal domains, which makes it difficult to observe and is usually a missing part in the solved structures of most of these protein.

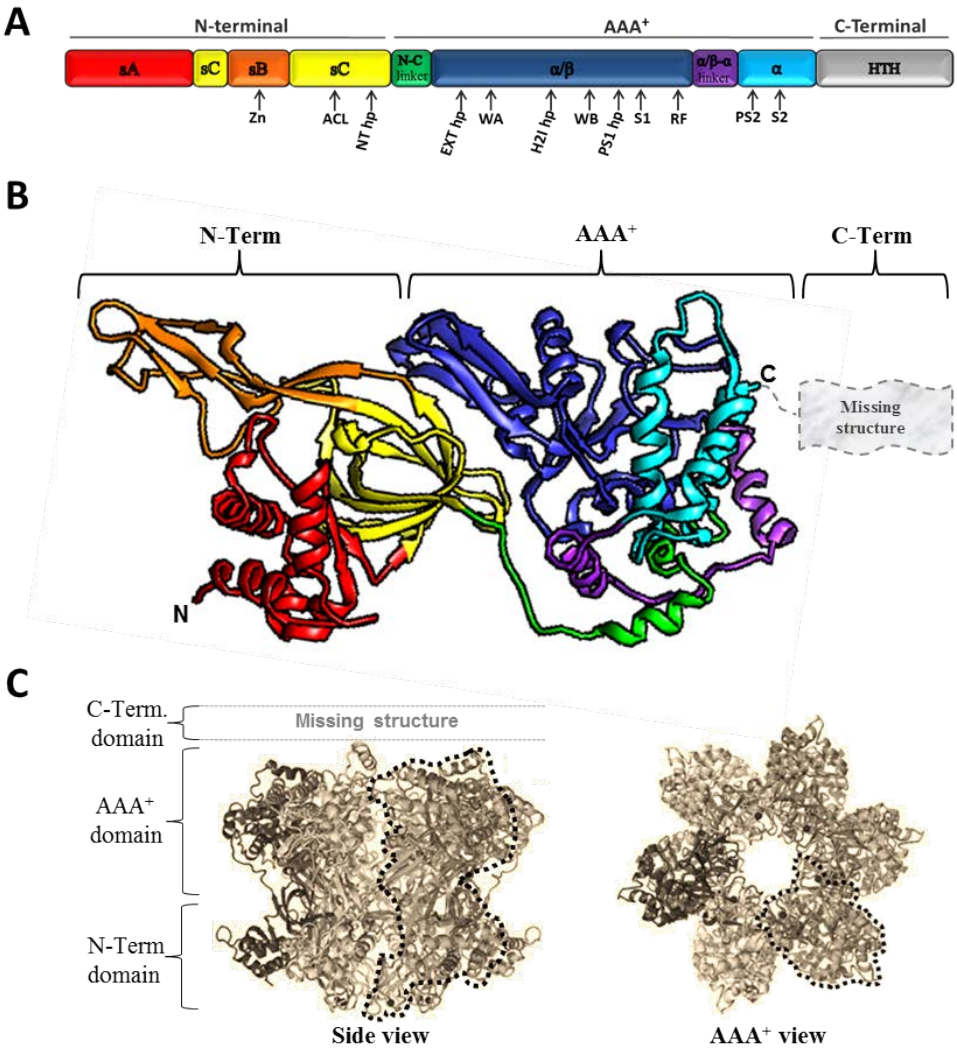


Figure 7. General structural features of MCM.(A) Linear schematic representation of the structural motifs of *Sso*MCM. Structural features are labeled and colored as follows: **sA** (red box), subdomain A; **sB** (orange box), subdomain B; **sC** (yellow), subdomain C; **N-C linker** (green box), amino and carboxyl domains linker; **α/β** (blue box), α/β domain; **α/β-α linker** (purple box), linker between domains α/β and α; **α** (cyan box), α domain; **HTH** (grey box), C-terminal domain helix-turn-helix motif; **Zn**, Zinc finger; **ACL**, allosteric communication loop; **NT hp**, N-terminal β-hairpin; **Ext hp**, external β-hairpin; **WA**, Walker A motif; **H2I hp**, helix 2 insert β-hairpin; **WB**, Walker B motif; **PS1 hp**, presensor 1 β-hairpin; **S1**, sensor 1; **RF**, arginine finger motif; **PS2**, presensor 2 insertion; **S2**, sensor 2. **(B)** Ribbon diagram of the near full-length structure of *Sso*MCM (PDB-3F9V). Domains and linkers are colored with the same code-color used in panel **A**. **(C)** Side and top views of a hexameric atomic model of *Sso*MCM lacking the C-terminal domain. Black dotted lines contour one of the monomers. (Figures of panel **C** were adapted from (Brewster et al. 2008)).

MCM proteins can form ring-shaped complexes (Figure 7C) with ATPase and helicase activities, showing a DNA unwinding and translocation polarity in the 3´→5´ direction (reviewed in (Costa & Onesti 2009), (Bochman & Schwacha 2009), (Brewster & Chen 2010), (Bell & Botchan 2013)). Different models have been proposed to explain how MCM effects the DNA unwinding (Brewster & Chen 2010), been the steric exclusion model and the side-channel extrusion model, the most

accepted ones (Bochman & Schwacha 2009) (Figure 8). Fluorescence resonance energy transfer (FRET) studies in archaea *Sso*MCMs support a steric exclusion model in which one strand of DNA (3'-tail) go all along the helicase central channel, while the second strand (5'-tail) is displaced out ahead by the helicase (McGeoch et al. 2005), (Rothenberg et al. 2007), albeit maintaining a weak and dynamic interaction with the helicase surface (Graham et al. 2011). Less clear is the helicase mechanism of eukaryotic MCM complexes, and the actual path that follows the 5'-tail DNA strand remains to be clarified.

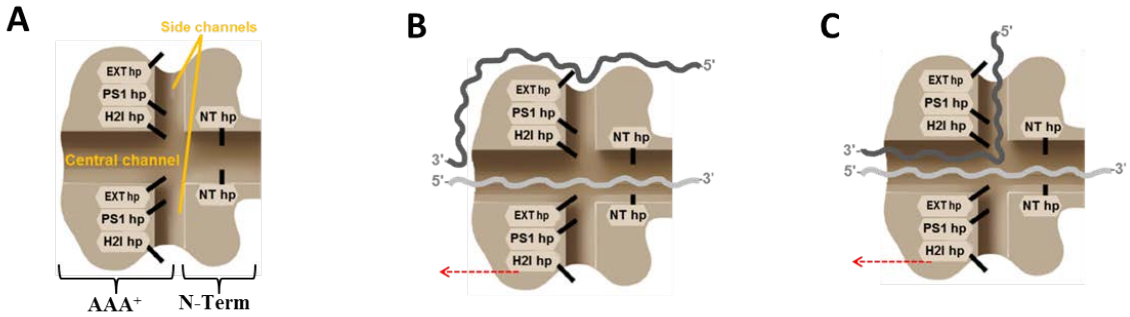


Figure 8. Proposed DNA unwinding modes by MCM helicase.(A) Schematic representation of a MCM hexameric helicase, showing a longitudinal cut, parallel to the central channel. The respective labeled β-hairpins are represented by short black bars. **(B)** Steric exclusion model and **(C)** side-channel extrusion model, proposed to explain MCM´s helicase activity. Light gray and dark grey wavy lines represent the DNA leading-strand and lagging-strand templates, respectively. Red dashed arrows indicate the movement directionality of the helicase (Figures adapted from (Brewster et al. 2008)).

Mostly found as hexameric single or double rings, the asymmetries and conformational changes displayed by some MCM helicases suggest a certain degree in both intra- and inter-domain flexibility (Costa et al. 2006a). Recently, the structure of a full-length *Mth*MCM dodecameric complex was solved by using a combination of small-angle neutron scattering (SANS) and all-atom molecular modeling techniques, where the free movement of the AAA+ core domains could be observed. The flexible loop connecting the AAA+ domain with the WH motif suggest that the C-teminus is very flexible (Krueger et al. 2014). The observed asymmetric positioning between the two tiers of the MCM2-7 ring in the EM structure of the *Dme*CMG-DNA-ATPγS complex, reveal that the N-terminus and AAA^+ regions can behave conformationally independent of each other (Costa et al. 2014).

Similar to other AAA^+ proteins, the binding of ATP to MCM helicases and its subsequent hydrolysis is thought to be carried out at the interface between two consecutive protomers (Figure 9). One of

the monomers provides the motifs for ATP binding in *cis* (Walker A, -WA-) and positioning of the nucleophilic water molecule (Walker B -WB- and sensor 1 -S1-), while the adjacent monomer contributes with the *trans* motifs (arginine finger -RF- and sensor 2 -S2-) that makes contact with the γ-phosphate of the same nucleotide. (Iyer et al. 2004), (Erzberger & Berger 2006), (Brewster et al. 2008). In this way, the RF from one subunit catalyzes the hydrolysis of the ATP bound to the next subunit. The way S2 function in other AAA+ proteins is in *cis*, but in MCM proteins the insertion of presensor 2 (PS2) allows S2 to act in *trans* (Brewster et al. 2008), (Erzberger & Berger 2006). The sequence of PS2 notably diverges among the six eukaryotic MCM proteins (Bochman & Schwacha 2009). A conserved region, called the allosteric communication loop (ACL), facilitates the coordination of the N- and C-terminal motifs of adjacent subunits that are involved in the binding and hydrolysis of ATP (Barry et al. 2009). Near the N-terminus of the protein, there is a zinc finger motif (Zn finger) that is involved in the folding stabilization of the N-terminal domain. Another remarkable feature is the presence of four β-hairpins that protrude from the surface (Brewster et al. 2008). Starting from the amino terminus of the sequence, the first hairpin (hp) is the N-terminal hp (NT-hp), followed by the exterior-hp (EXT-hp), the helix 2 insert hp (H2I-hp) and, finally, the presensor 1 hp (PS1-hp).The β-hairpins are involved in DNA interactions (Brewster & Chen 2010), (Bell & Botchan 2013).

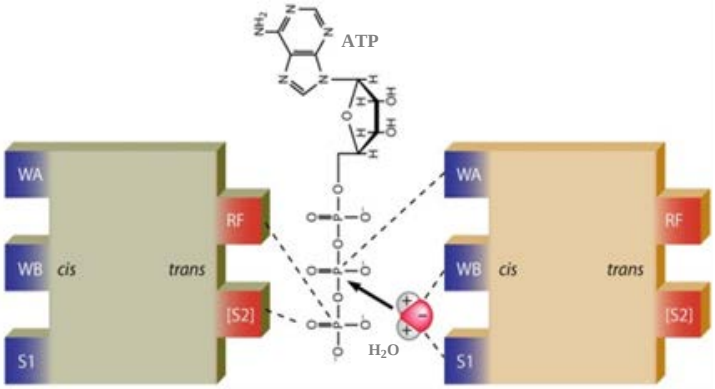


Figure 9. Active sites in AAA$^+$ domains are formed at the interface between adjacent protomers. In MCM proteins, the motifs WA, S1 and WB act in *cis*, on the other hand RF and S2 motifs act in *trans* to hydrolyze an ATP molecule (Figure adapted from (Bochman & Schwacha 2009)).

1.4.3. The eukaryotic MCMs

In eukaryotes, all six MCM paralogous proteins, MCM2 to MCM7 (MCMs 2-7), are required for replication of DNA. The heterohexameric MCM2-7 helicase has been purified isolated or as part of bigger complexes (the pre-replicative complex, pre-RC (Sun et al. 2014); the intermediate ORC-Cdc6-Cdt1_MCM2-7 complex, OCCM (Sun et al. 2013); or as part of the holo-helicase complex Cdc45-(MCM2-7)-GINS, CMG (Costa et al. 2011), (Costa et al. 2014)). The previously mentioned evidences pointed to the MCM2-7 heterohexamer as the replicative helicase in eukaryotes (Chong et al. 1996), (Kearsey & Labib 1998), (Kearsey et al. 1996), (Mcm 1994). The activation of the MCM2-7 helicase seems to depend on its phosphorylation by DDK and CDK kinases (Remus & Diffley 2009), followed by its association with GINS and Cdc45 (Ilves et al. 2010). *In vitro* helicase activity of the isolated MCM2-7 complex has been observed in budding yeast and, more recently, in human, but only under specific buffer anion conditions (requiring the addition of glutamate and/or acetate) that seems to induce the closure of the MCM2/5 gate (Bochman & Schwacha 2008), (Hesketh et al. 2015).

The subunit organization within the MCM2-7 ring complex was experimentally established as MCM5→3→7→4→6→2 (within each interacting pair, the subunits to the left of the arrow provide the motifs in *trans*, while the subunits to the right contribute with the motifs in *cis*) (Ishimi 1997b), (Schwacha & Bell 2001), (Davey et al. 2003), (Costa et al. 2011) (Sun et al. 2013), (Sun et al. 2014), where the interaction of MCM2 with MCM5 forms the close/opening ring interface that has been observed as a gap by EM (Figure 10A) (Costa et al. 2011), (Lyubimov et al. 2012). This gap could work as an ATP-modulated gate for ssDNA to pass through, so it can access the central channel of the ring (Bochman & Schwacha 2008), (Costa et al. 2011), (Sun et al. 2013).

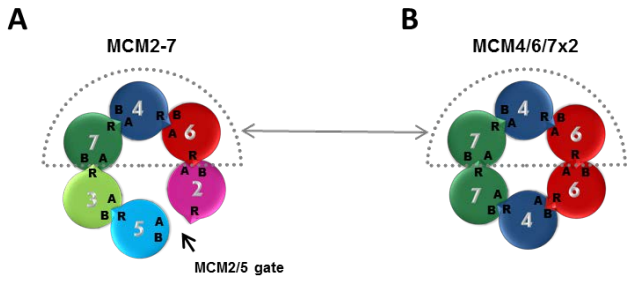


Figure 10. Architecture of MCM helicases. Diagrams for the protomer organization of **(A)** MCM2-7 and **(B)** the proposed putative protomer arrangement in the MCM4/6/7x2 hexamer, based on the known physical interactions between the different MCM proteins. Grey doted semicircle figure highlights the MCM trimer composed by MCM4, 6 and 7, which would putatively share a similar organization in both helicases. The position of the motifs: Walker A, Walker B, and arginine finger, are indicated by capital letters "A", "B", and "R", respectively.

In addition to the MCM2-7 heterohexamer, it has been found eukaryote MCMs proteins can be associated in several other combinations, giving rise to dimers (Tye 1996), (Burkhart et al. 1995), (Sherman & Forsburg 1998), trimers (Musahl et al. 1995), tetramers (Ichinose 1996), hexamers (Tye 1996), (Richter & Knippers 1997) and double-hexamers (Evrin et al. 2009), (Remus et al. 2009). Some of the MCM polypeptides are also able to self-interact (Tye 1996), (Yabuta et al. 2003), (Tran et al. 2010). Deletion and mutational studies were carried out to better understand the role of each of the six MCMs proteins within the different subcomplexes. It was shown that each of the six types of MCM proteins contributes differentially to the biophysical and biochemical properties of the complex, involving the interaction with other different proteins/complexes (You & Masai 2008), (Wei et al. 2010), (Liu et al. 2012), (Costa et al. 2011), (Sun et al. 2013), (Nguyen et al. 2012), oligomerization and stabilization of the hexameric complex, ATP hydrolysis and, DNA binding and helicase activity (Lee & Hurwitz 2000), (Ma et al. 2010), (Xu et al. 2013), (Schwacha & Bell 2001), (Bochman et al. 2008), (Davey et al. 2003), (Bochman & Schwacha 2007), (Ying & Gautier 2005). From the above, it was concluded that the MCM4/6/7 heterotrimer forms the catalytic core of the MCM2-7 helicase, whereas MCM2, 3 and 5 play a negative regulatory role (Ishimi 1997b), (Lee & Hurwitz 2000), (Sato et al. 2000), (Schwacha & Bell 2001). In agreement with all these results, the different oligomers present particular properties and levels of activity.

1.4.4. MCM4/6/7

Aside from the MCM2-7 helicase, significant effort has been invested in the study of the biochemical properties of the MCM4/6/7 helicase, an hexameric ring complex formed by two MCM4/6/7 heterotrimers (hereinafter alternatively referred to as MCM4/6/7x2) which, in contrast to the MCM2-7 hexamer, shows an intrinsically helicase activity in a persistent manner (You et al. 1999). MCM4/6/7x2 presents limited processivity (Ishimi 1997b), (Lee & Hurwitz 2000). (Lee & Hurwitz 2001), but a double heterohexameric complex (a dimer of MCM4/6/7x2) that assemble on forked DNA substrate, has been shown to be a processive helicase (Lee & Hurwitz 2001). MCM4/6/7x2 has been isolated from diverse systems, and both its ATPase and DNA helicase activities are conserved across species (Sherman & Forsburg 1998), (Prokhorova & Blow 2000), (Holthoff 1998), (Davey et al. 2003), (Coué et al. 1998), (Ichinose 1996), (Ishimi 1997a), (You et al. 1999), (Lee & Hurwitz 2000), (Sato et al. 2000), (Lee & Hurwitz 2001), (You et al. 2002), (Biswas-

Fiss et al. 2005), (You & Masai 2005), (Bochman & Schwacha 2007), (You & Masai 2008), (Xu et al. 2013), suggesting that this complex may be implicated in important cellular processes. Although currently the biological role of MCM4/6/7x2 remains unknown, it is not surprising that dysregulation or mutations of its components are associated with several clinical conditions, including cancer (Watanabe et al. 2012), (Gineau et al. 2012), (Hughes et al. 2012), (T. Zheng et al. 2014).

Regarding the catalytic role of the MCM4/6/7 trimer subcomplex and the regulatory role of the MCM2, 3 and 5 proteins, it was observed that MCM4/6/7x2 showed both a better dsDNA binding performance and a higher ssDNA association rate when compared to MCM2-7 (Bochman & Schwacha 2007). The authors of that work proposed that these differences are a consequence of conformational changes in a putative regulatory site in the MCM2/5 gate. Whether there could exist a similar gate in MCM4/6/7x2 remains an important question to be answer. Another observation also supporting the idea of the regulatory action of MCM2, 3 and 5, was that incubation of either MCM2 or MCM3/5 with MCM4/6/7x2, causes a disassembly of the dimeric complex of MCM4/6/7 along with the loss of DNA helicase activity (Lee & Hurwitz 2000).

All protomers surely contribute in some general way to the proper function of the helicase but some principal roles have been associated to each of the three types of subunits, MCM4, 6 and 7. Phylogenetic analysis suggest that MCM4 is the most ancient form of eukaryotic MCMs (Kearsey & Labib 1998). WA and WB motifs of MCM4 play key roles in the ssDNA binding activity of the helicase (Bochman & Schwacha 2007), (Gómez et al. 2002), (You et al. 1999). It has been suggested that in fission yeast, the C-terminus of MCM4 blocks the DNA unwinding activity of MCM2-7 once there is enough ssDNA at the stalled replication fork for checkpoint activation. On the other hand, the MCM4 C-terminal part is also required for efficient resumption of fork progression during recovery from the replication block (Nitani et al. 2008). Mutations in the Zn-finger of MCM4 promote the dissociation of MCM4/6/7x2 into its trimers, so it seems that dimerization of MCM4/6/7 is somehow directed by the Zn-finger of the MCM4 subunits. The above observation suggest that MCM4 should presents at least two different domains with particular oligomerization roles, one for the formation of the MCM4/6/7 trimer and the other, within the N-terminus, involved in the dimerization of the trimers themselves (You et al. 2002). In turn, MCM6 plays a major role in binding of ATP, and MCM7 is required for hydrolysis of ATP and helicase activity (You et al. 1999). Based on these considerations, a simplified model for the DNA

unwinding mechanism of MCM4/6/7x2 was proposed (You et al. 2002), in which there is a sequential participation of the protomers and each protein class (e.i. MCM4, 6 or 7) makes a major-contribution to a different task. First, DNA binds to the helicase by interacting with MCM4, then, an ATP molecule binds to MCM6 causing a conformational change in the complex that would facilitate the subsequent hydrolysis of the nucleotide by MCM7. A tight communication and coordinated work between all subunits in MCM4/6/7x2 is essential in order to fully carry out its helicase activity.

Usually isolated as the hexameric complex MCM4/6/7x2 (Ishimi 1997a), (Lee & Hurwitz 2000), (Yabuta et al. 2003), (Ishimi 1997b), (You et al. 1999), (Ishimi 1997a), (Sato et al. 2000), (Lee & Hurwitz 2000), (Ma et al. 2010), (Su et al. 1996), proteins MCM4, MCM2 and MCM7 have been also purified as an stable single MCM4/6/7 heterotrimer (Musahl et al. 1995), (Sherman et al. 1998), (Ma et al. 2010), (Ichinose 1996), (Ishimi 1997a). Although it is not clear which are the factors directing the oligomerization state of MCM4/6/7, a study with *S.cerevisiae* proteins, in which MCM4, 6 and 7 were individually purified and then mixed together, showed a shift from the trimer to the hexamer form after incubation with ATP, a cofactor required for helicase activity. Incubation with ATPγS had a similar oligomerization effect to ATP, and it was also found that ADP fosters the formation of the hexamers but to a lesser extent (Ma et al. 2010). In agreement with the idea that the active complex is an hexamer and not a single trimer, a mutational study showed that the RF motifs of the three proteins are required for a full ATPase activity (Ma et al. 2010).

Despite certain disagreement in the literature regarding some of the direct pair interactions between eukaryote MCM subunits (including some self-interactions), most of the experimental reports support an architecture model for the MCM4/6/7x2 hexameric ring in which the two MCM4/6/7 trimers face each other in a way that the two MCM6 subunits binds together (MCM6-MCM6) as well as the two MCM7 proteins (MCM7-MCM7), and each of the MCM4 is located between the two different homotypic pairs, that is, MCM4-6-6-4-7-7 (Yabuta et al. 2003), (Yu et al. 2004), (Xu et al. 2013), (Figure 10. B). To the best of our knowledge, there is only one study where an alternative model is proposed, suggesting a direct binding of MCM6 and MCM7 (Ma et al. 2010).

To date there are plenty of studies focused on the biochemical properties of the hexameric helicase MCM4/6/7x2. Contrarily, at the structural level there are only a few works showing some

2D images of the closed ring-shaped complex, so many questions remain yet to be answered. In order to better understand the structural and functional organization of MCM4/6/7x2, we have taken advantage of the similarities that exist between this helicase and MCM2-7.

In this work we present the structural study of diverse MCM4/6/7 complexes as an important contribution to continue with the process of elucidating the functional mechanism of this macromolecular motor. Our data suggests how MCM4/6/7 trimers dimerize, and we propose, based on a number of novel and revealing conformational changes of the hexameric complex, a sequence that would give rise to a putatively active MCM4/6/7x2 structure, able to perform DNA binding and helicase activities.

1.5. Biological system: TWINKLE, the human mitochondrial DNA helicase

Twinkle, the mitochondrial replicative DNA helicase, is a ring-shaped enzyme (Ziebarth et al. 2010), (Fernández-Millán et al. 2015) required for strand separation at the replication fork of mitochondrial DNA (mtDNA), making it essential for mtDNA maintenance and regulation of its copy number (Tyynismaa et al. 2004). The gene C10orf2 on chromosome 10q24 encoding the Twinkle protein was originally identified through its link with a neuromuscular disorder associated with mtDNA damage (Spelbrink et al. 2001), a clinical syndrome known as autosomal dominant progressive external ophthalmoplegia (adPEO), which is associated with mutations in the Twinkle gene (Fratter et al. 2010) and multiple mtDNA deletions (Suomalainen et al. 1997).

Mitochondria are popularly known as the “powerhouse of the cell” because they are responsible for producing most of the ATP used as source of chemical energy for the cells (Scheffler 2007), which requires an intact and functional mitochondrial genome. These organelles had also been found to be tightly involved in the execution of diverse cellular events, such as the regulation of metabolism, cell-cycle control, development, antiviral responses and cell death (Mcbride & Neuspiel 2006), (Tower 2015). Cell´s homeostasis and adaptability depend on the qualitative and quantitative status of the mitochondrial genome. Mitochondria contain their own genome. Normally, mtDNA is organized as a circular, covalently closed, double-stranded DNA (dsDNA), which is thought to be derived from the α-proteobacterium that eventually became the mitochondrion (Gray et al. 1999), (Burger et al. 2003). The human mtDNA molecule (and that of mammals in general) is around 15.5 kb, and its two strands are differentiated by their nucleotide content, with a guanine-rich strand referred to as the heavy strand (H-strand) and a cytosine-rich strand referred to as the light strand (L-strand). Together, both H-strand and L-strand contain 37 genes (Figure 11), all of them essential for mitochondrial function (Kyriakouli et al. 2008). Alterations in the mtDNA molecule or its copy number are associated with a number of clinical conditions (Mao & Holt 2009), (Nogueira et al. 2014), (Spelbrink et al. 2001), (Suomalainen et al. 1997).

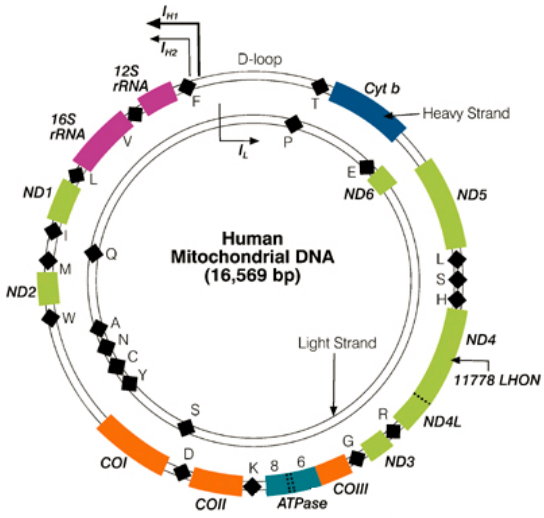


Figure 11. The human mitochondrial genome. This small (16,569 bp) genome is almost completely transcribed from both strands, initiating from either one of two promoters (IH1, IH2) on the Heavy (H-) strand, or the single promoter (IL) on the Light (L-) strand. All of these promoters and elements involved in replication initiation are found in the displacement (D-) loop, the only major non-coding region. The genome encodes 22 mt-tRNA (black diamonds), 2 mt-rRNA genes (fuscia) and 13 protein-coding genes (olive, ND1-6 encoding members of NADH:ubiquinone oxido-reductase; blue, Cytb encoding apocytochrome *b* of ubiquinol:cytochrome *c* oxido-reductase; orange, COI-III encoding members of cytochrome *c* oxidase: aquamarine, ATPase 6, 8 encoding two members of Fo-F1 ATP synthase). Single letter code is given for each mt-tRNA-encoding gene (Figure and legend modified from (Kyriakouli et al. 2008)).

1.5.1. The mtDNA replisome

The proteins that constitute the minimal mtDNA replisome, a protein complex required for the maintenance of mtDNA (Korhonen et al. 2004), (Spelbrink et al. 2001), (Wanrooij & Falkenberg 2010), (Tyynismaa et al. 2004), are all encoded by nuclear genes and form a small set that includes Twinkle helicase, both the catalytic (α) and accessory (β) subunits of DNA polymerase γ (pol γ), the mitochondrial single-stranded DNA-binding protein (mtSSB) (Korhonen et al. 2004), and the mitochondrial RNA polymerase (mtRNApol), which functions as the mtDNA primase (Wanrooij et al. 2010), (Shi et al. 2008), (Reyes et al. 2013), (Pham et al. 2006), (Shi et al. 2008). Interestingly, it seems that frequently along mitochondrial history, functional redirection of ancestral proteins has happened, and such is the case for mtDNA replisome components, where Twinkle, the catalytic subunit pol γ-α and mtRNApol share ancestry with the gp4 primase-helicase, the gp5 DNA polymerase and the gp1 RNA polymerase from the bacteriophage T7, respectively; whereas mtSSB is a homologue of the homotetrameric SSBs from eubacteria (Shutt & Gray 2006a), and the accessory subunit pol γ-β evolved from the class II aminoacyl-tRNA synthetases of eubacteria (Uhn 1999), (Fan et al. 2006).

1.5.2. Twinkle architecture: Known and proposed functions

Sequence similarities between the mitochondrial replicative DNA helicase and the bifunctional primase-helicase of bacteriophage T7 gen 4 (T7 gp4) (Spelbrink et al. 2001), (Shutt & Gray 2006b), together with additional studies showing related physical features of their architectures (Ziebarth et al. 2007), assigned Twinkle to the superfamily 4 (SF4) of helicases. Both proteins present a general structure where a flexible linker connects distinct N-terminal and C-terminal domains (Figure 12).

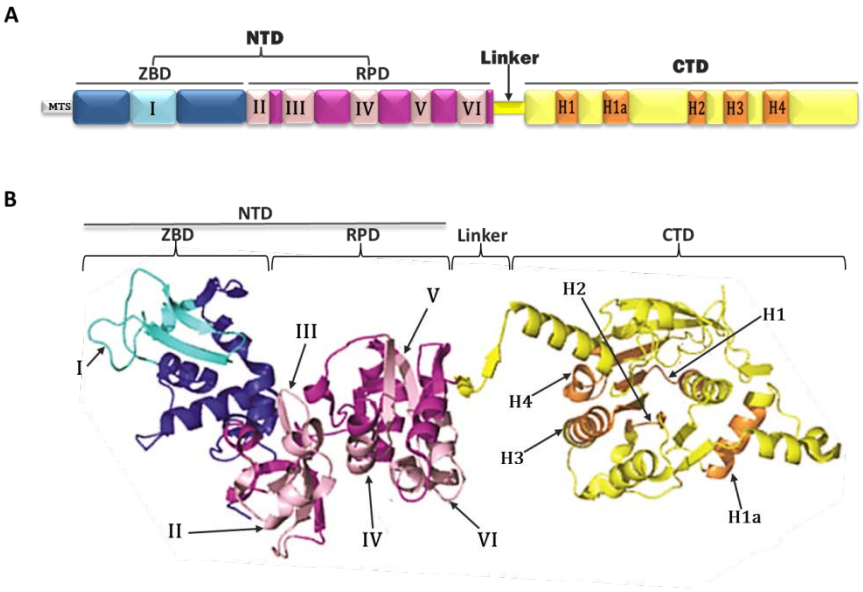


Figure 12. General structural features of the replicative mtDNA helicase Twinkle.(A) Linear schematic representation of the structural motifs, which are labeled and colored as follows: The N-terminal domain (NTD) is formed by two smaller domains, the zinc binding domain (ZBD, in cyan and blue) and the RNA polymerase domain (RPD, in pink and fuchsia); Motifs from the NTD (I, II, III, IV, V and VI) and those from the CTD (H1, H1a, H2, H3 and H4). (B) Ribbon diagram of an homology model of Twinkle, showing the approximate localization of the structural motifs, where the colors and nomenclature showed in panel A are maintained (Figure from panel B has been adapted from (Fernández-Millán et al. 2015)).

Twinkle has a NTPase-dependent unwinding activity with a 5´to 3`translocation directionality (Korhonen et al. 2003), as well as a tendency to oligomerize into toroidal complexes (Ziebarth et al. 2010), (Fernández-Millán et al. 2015), which are features showed by the SF4 members (Singleton et al. 2007). While it is true that Twinkle can oligomerize in a NTP-independent way, its oligomeric status (hexamerization and/or heptamerization) is affected by different ionic conditions and cofactors (Ziebarth et al. 2010), (Farge et al. 2008), (Fernández-Millán et al. 2015); for

example, it has been observed that high salt conditions (Ziebarth et al. 2010), (Fernández-Millán et al. 2015) or addition of MgUTP (Sen et al. 2012) favors the presence of hexamers over higher oligomers. In a similar way, T7 gp4 forms both hexamers and heptamers (Toth et al. 2003), (Crampton et al. 2006), where the first oligomeric state predominates in presence of triphosphate, while the use of diphosphate produces a higher proportion of heptamers over hexamers (Crampton et al. 2006). In the presence of DNA and nucleotide (both dTTP and dTDP), T7 gp4 is detected as hexamers that bind ssDNA. On the other hand, once formed, T7 gp4 heptamers are not able to efficiently bind either ssDNA or dsDNA (Crampton et al. 2006). Twinkle interacts with a variety of DNA substrates, even in the absence of cofactors, including both linear and circular ssDNA and linear dsDNA (Jemt et al. 2011), and shows higher binding affinity for dsDNA than for ssDNA (Sen et al. 2012), (Ziebarth et al. 2010), (Farge et al. 2008),(Longley et al. 2010), while T7 gp4 binds preferably to ssDNA over dsDNA, requiring cofactors for binding ssDNA (Matson et al. 1985), (Hingorani & Patel 1993). Currently, three different DNA binding sites in Twinkle have been detected; one for dsDNA and two for ssDNA, where the latter sites are putatively localized one in the central channel and the other on the external surface of the helicase ring (Sen et al. 2012). Similarly to T7gp4, Twinkle requires substrates with both a single-stranded 5'-DNA loading site and a short 3'-tail (which resembles the structure of a replication fork) to efficiently initiate the unwinding of the DNA duplex (Korhonen et al. 2003)(Patel & Picha 2000).

The C-terminal domain (CTD) of Twinkle shows high homology with the equivalent region in T7gp4 (Spelbrink et al. 2001), (Shutt & Gray 2006b), corresponding in both cases to their helicase domain and an important part for oligomerization of these enzymes, where the conserved catalytic motifs Walker A and Walker B (within SF4 are known as motifs H1 and H2, respectively) found in helicases and translocases are housed (Singleton et al. 2007). The three additional motifs H1a, H3 and H4 present at Twinkle´s CTD, are characteristic signatures of SF4 (Singleton et al. 2007) (Figure 12). In Twinkle, the linker region is important for oligomerization (Guo 1999), and several adPEO disease-related mutations reside within this region (Ji et al. 2014), (Dündar et al. 2012). Regarding the N-terminal domain (NTD) of the mtDNA helicase, in mammals, this region of the protein has diverged dramatically from that of other metazoans, losing most of the residues present in T7gp4 that are essential for the primase activity so that, in mammals, Twinkle doesn´t show this function (Shutt & Gray 2006b). In particular, human Twinkle lacks three of the four conserved cysteine residues found in motif I of T7 gp4 (Spelbrink et al. 2001), (Shutt & Gray 2006b), a zinc-finger motif crucial for DNA binding and primase activity (Kusakabe & Richardson 1996), (Ilyina et al. 1992) (Figure 13).

However, primase motifs II to VI are present and are well conserved in Twinkle, with the exception of important residues involved in binding Mg^{2+} which are missing (Spelbrink et al. 2001), (Shutt & Gray 2006b).

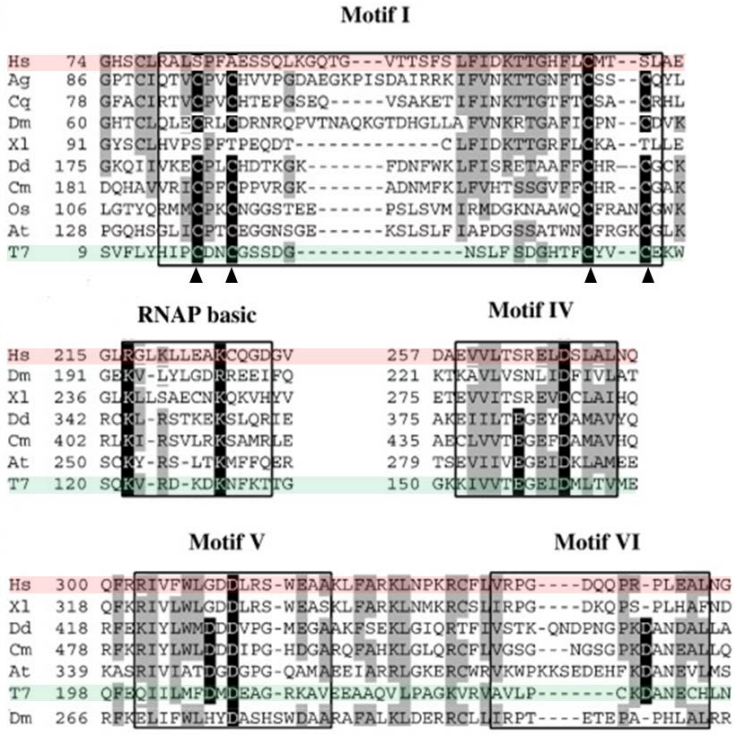


Figure 13. Sequence alignment of the NTD between representative Twinkle homologues and T7 gp4. Conserved sequence motifs are highlighted into black boxes and labeled with the respective name. Critical residues for T7 gp4 and those that are conserved in mtDNA helicases are indicated in black and gray, respectively. Organism´s name abbreviations: man (Hs, highlighted in red), fly (Dm), African malaria mosquito (Ag), Southern house mosquito (Cq), frog (Xl), amoeba (Dd), red alga (Cm), rice (Os), mouse-ear cress (At) and bacteriophage T7 gp4 (T7, highlighted in green). Black arrowheads point to four conserved cysteine residues in motif I of T7 gp4. (Figure and legend modified from (Matsushima & Kaguny S. 2009)).

The NTD of Twinkle is classified into the group of prokaryotic DnaG-type primases (Spelbrink et al. 2001), (Shutt & Gray 2006b), by its predicted fold showing an overall Toprin (topoisomerase-primase) architecture (Aravind et al. 1998), even though the enzyme lacks primase activity. Members of DnaG family present a NTD region with a zinc binding domain (ZBD) toward the N-terminal end, followed by a short linker connected to a RNA polymerase domain (RPD) (Haven 1994), (Tougu et al. 1994). The RPD can vary among primases but they still share in common the Toprim subdomain, which in T7 gp4 and in DnaG is required for catalyzing primer elongation (Kato et al. 2003),(Keck et al. 2000), (Podobnik et al. 2000). The ZBD of the primase binds to ssDNA and

recognizes specific nucleotide triplet patterns from priming sites in the ssDNA template (Lee et al. 2012), which subsequently passes to the RPD, where binding and the catalytic condensation of NTPs occurs; RPD is required for the template-directed coupling of the NTPs (Kuchta & Stengel 2010).

Twinkle mutants lacking its NTD can only support the synthesis of short DNA products (~ 1 kb) (Farge et al. 2008), so despite the fact that the NTD of Twinkle is not absolutely required to carry out the most basic function of the mtDNA replisome, it is important for the efficient ssDNA binding and dsDNA unwinding activities of the enzyme, which are required for formation of long DNA products (Farge et al. 2008). Furthermore, it seems that the NTD also contributes to the proper oligomerization of the helicase ring complex (Farge et al. 2008). Other authors have proposed possible additional roles of the NTD, such as stimulation of primer utilization (Shutt & Gray 2006b). Removal of ZBD in human Twinkle, produces a decrement in ssDNA binding compared with that of the full-length protein, affecting also the ssDNA-stimulation of the ATPase activity of the enzyme (Farge et al. 2008).

On the other hand, the lack of the ZBD did not affect the protein binding to dsDNA, suggesting that the ZBD contributes specifically to ssDNA binding. It has been proposed that, in T7gp4, the initial contact of a DNA binding site at its primase domain with a DNA molecule, would produce a transient opening of the ring allowing the access of DNA to the central channel, after which the gap is closed again (Ahnert et al. 2000), (Farge et al. 2008), (O'Shea, V.L. and Berger 2014). An alternative mechanism for DNA loading of T7 gp4 has been proposed based on the observation that it forms both hexameric and heptameric complexes (Patelt 1995), (Toth et al. 2003), (Crampton et al. 2006). It is suggested that T7 gp4 heptamers could lose one protomer, which would lead to an opening in the resulting hexamer (Crampton et al. 2006). Since Twinkle forms both hexameric and heptameric toroidal complexes (Ziebarth et al. 2010), (Fernández-Millán et al. 2015) and is able to load by its own on the DNA, similar mechanism to the ones proposed for T7 gp4 could be also considered in the case of Twinkle as possible strategies to create a gap that allows the helicase ring to load onto DNA without assistance from additional factors. The assembly into different oligomeric ring-shaped complexes is a feature showed by a number of AAA^+ ATPases, including some members of superfamily 6 (SF6) of helicases, as for example MCM helicases from different archaea and eukaryotic organisms (Yu et al. 2002), (Fletcher et al. 2003), (Adachi et al. 1997), (Lee & Hurwitz 2001).

Interestingly, in addition to the already known NTPase-dependent helicase activity of Twinkle, the antagonistic function, an efficient ability of annealing two complementary ssDNAs, has been detected to occur both with and without cofactors (Sen et al. 2012). Indeed, it was observed that under certain conditions (absence of gp2.5 or ssDNA trap) the annealing activity competes with the unwinding activity. Twinkle´s annealing activity is sensitive to the ionic strength of the environment, being inhibited at concentrations higher than 150mM NaCl (Sen et al. 2012), which can be partially explained by the fact that protein-DNA interactions start being disrupted at high salt conditions, and are abolished at 420mM NaCl (Yakovchuk et al. 2006).

Recently, the 3D structure of a hexameric complex of Twinkle that was isolated under high salt conditions (1M NaCl) and stabilized by chemical fixation with the GraFix method, was solved by cryo-electron microscopy (EM) (the density map is still not available on the Electron Microscopy Data Bank, EMDB) (Fernández-Millán et al. 2015) (Figure 14A). The 3D map shows a compact, two-tier ring conformation, which resembles the general shape of a T7 gp4 heptamer complex which naturally lacks the ZBDs (PDB-1Q57) (Toth et al. 2003) (Figure 14B).

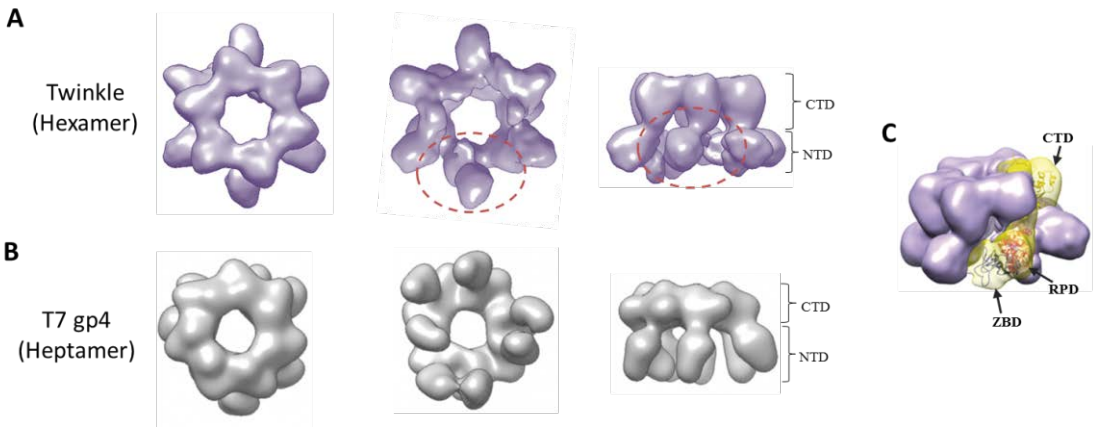


Figure 14. Ring-shape helicase 3D structures.(A) EM map of a Twinkle hexamer (Fernández-Millán et al. 2015) and (B) the structure of an heptamer of T7 gp4 (PDB-1Q57, filtered at 30Å). From left to right the CTD-, NTD- and lateral-views are shown, respectively. A red dashed line highlights a particular NTD in Twinkle ´s structure that is isolated from the rest of NTDs, which are interconnected through contacts with the ZBD of the neighbouring monomer.(C) Tilted view from Twinkle EM map, with one segmented subunit and a homology model fitted in (Figure from panel C was adapted from (Fernández-Millán et al. 2015).

The proposed flexible fitting of an atomic homology model of the full-length Twinkle protein (a chimeric model based on predicted secondary structure similarities between Twinkle and both the CTD of T7 gp4 and the NTD of DnaG) into the cryo-EM map (Figure 14C), accommodates the CTDs into the density of a closed six fold symmetry (C6) ring, showing similar interfaces than the ones observed in both the T7 gp4 hexamer (PDB-1E0J and 1E0K) (Singleton et al. 2000) and heptamer (PDB-1Q57) structures. The RPDs are fitted in the second floor of the map, which unlike the CTDs layer, shows a slightly asymmetric ring. Four smaller extra densities, located in a level slightly lower than the RPD layer, are assigned to some of the ZBDs, being asymmetrically positioned in a way that interconnects five of the RPDs, while the last RPD remains disconnected from the rest. The lack of two ZBDs in the density map is attributed to a higher flexibility of these domains. The different interface contacts within this hexamer of Twinkle are: *in cis,* ZBD/RPD and *in trans*, both ZBD/RPD and RPD/CTD. The above is in agreement with studies on the related SF4 proteins DnaG (Corn et al. 2005) and T7 gp4 (Lee & Richardson 2002), where ZBD/RPD interactions, both *in cis* and *in trans*, have been detected; additionally, contacts of RPD and RPD-CTD linker with the CTD of the neighbouring protomer have been observed in T7 gp4 (Singleton et al. 2000). In the same study of Twinkle protein (Fernández-Millán et al. 2015), negative staining (NS) EM examination of the high-salt, unfixed sample, showed both hexamers and heptamers with a radial arrangement of the subunits where a central pore was formed by the interacting edge of the protomers, while the opposite side of all the subunits remained extended to the outside with variable orientations. Small angle X-ray scattering (SAXS) analysis of the particles in solution at 1.5M NaCl, showed a dynamic sample with a mixture of extended conformations of both hexamers and heptamers (with a calculated ratio 3/2), where the protein presented high flexibility at the inter-domain connectors.

Aside from observing that Twinkle forms both flexible hexamers and heptamers, It is also known that low salt concentrations result in protein aggregation (Ziebarth et al. 2010), (Fernández-Millán et al. 2015), which have made difficult its structural analysis under more physiological, lower ionic strength conditions. Obviously, much more work is still required to elucidate the structural mechanisms that control the different functions of this essential enzyme for cell biogenesis. Here, we present an EM analysis of the human Twinkle, by comparing the full-length protein (mature Twinkle includes residues 43-684) with a truncate construct lacking the ZBD (ΔZBD protein includes residues 147-684), under salt concentrations between 250-330mM NaCl. Our comparative analysis between the full-length and N-terminal truncate proteins provides direct

visual evidence of the fundamental role of the ZBD on the conformational behavior of the ring complexes. Indeed, deletion of the ZBD significantly restricts the capability of the enzyme to acquire the whole range of additional conformational movements showed by the WT, which we suggest would be very likely required for the correct function and interactions that must be normally carried out by Twikle. We also report, for the first time, images of octameric and pentameric ring-like complexes of Twinkle which coexist in a small proportion with a majority population of the already reported hexameric and heptameric species. Our findings of structural details of Twinkle flexibility and the influence of the ZBD provide new insights to understand the mechanistic features displayed by the mitochondrial replicative DNA helicase in mammals.

Chapter 2

2. Objectives

The general aim of this thesis is the structural characterization by transmission electron microscopy of seceral macromolecular complexes, such as the centrosomal human protein construct CPAP$^{897-1338}$ and the mouse and the human helicases MCM4/6/7 and Twinkle, respectively, being the three of them, highly dynamic and flexible complexes, able to form different oligomeric assemblies. In particular, the following objectives have been pursued:

1. Purification of human CPAP$^{897-1338}$ and isolation of different oligomeric complexes formed by this protein. This task has been followed by the structural characterization and the three dimensional reconstructions of the most stable assemblies.

2. Three dimensional reconstruction of different oligomeric and conformational states of mouse MCM4/6/7, and fitting of atomic structures into the obtained density maps for the putative localization of structural domains.

3. Two dimensional characterization of the structure and flexibility of multiple oligomeric complexes formed by human Twinkle, determining the effect of the zinc binding domain on the structural dynamics of the helicase by comparing the full-length protein with a N-terminal deletion mutant. Additionally, initial low resolution three dimensional maps of the most representative oligomeric and conformational states have been obtained.

Chapter 3

3. Materials and methods

3.1. Obtaining the protein samples

3.1.1. CPAP$^{897\text{-}1338}$

3.1.1.1. Plasmid construct

The full-length human cDNA of CPAP (1338 residues) was kindly provided by Dr. Tang T. K., from the Institute of Biomedical Sciences, Academia Sinica. Taipei 115, Taiwan (Hung et al. 2000). The C-terminal residues 897 to 1338 of CPAP were amplified by PCR using the sense primer STO3203 that incorporates the BamHI restriction site (5'-GCGGATCCCCTGGTGACAATGCTCG -3'), and the STO3204 antisense primer containing BsrGI restriction site (5'-CGGTGTACATTACAGCTCCGTGTCCATTG -3'). To create a N-terminal 6xHis-tagged CPAP897-1338 recombinant DNA construct (His-CPAP897-1338), the amplified fragment was cloned into the BamHI-BsrGI site of the pST66Trc2-His expression vector (this is a pET3a modified plasmid created and kindly provided by Dr. Song Tan, form the Center for Gene Regulation, Pennsylvania State University, State College, PA. EE. UU.), which allowed the incorporation of a six histidine tag (His-tag) to the N-terminus of the CPAP897-1338 protein; in this way the recombinant expression plasmid pST66Tcr2-HisCPAP897-1338 was obtained. The strain E. coli DH5α was used as host for the cloning work. The isolation of the recombinant plasmid was performed using standard methods (Green & Sambrook 2012).

3.1.1.2. Protein expression

E. coli BL21(DE3)pLysS cells were transformed with the pST66Tcr2-HisCPAP$^{897\text{-}1338}$ recombinant plasmid, and protein expression was induced adding 0.2mM IPTG followed by an incubation at

23°C/16h. Solubility tests showed that CPAP$^{897-1338}$ protein is highly insoluble in low salt buffers, so it was purified and maintained most of the time in presence of 300mM NaCl, with the exception of a part of the protein sample that was dialyzed in buffer with 150mM NaCl to be posteriorly used for an ion exchange chromatography step. The cell pellet was resuspended in cold lysis buffer pH 7.4 (30mM Na_2HPO_4, 20mM NaH_2PO_4, 300mM NaCl, 10% v/v Glycerol, 0.1% Triton X-100, 25mM Imidazole, 10µg/ml DNase A, 10µg/ml RNase I, 0.5mg/ml Lysozyme, 1X Protease Inhibition Cocktail EDTA-free. Sigma) and lysed by French Press.

3.1.1.3. Protein purification by immobilized metal affinity chromatography (IMAC)

After centrifugation (40.000rpm/45min) of the cell lysate, the soluble His-tagged protein in the clarified supernatant was purified by immobilized metal affinity chromatography (IMAC) with a Talon resin (Talon superflow. GE Healthcare) following the manufacturer´s recommendations. The collected fractions were analyzed by SDS-PAGE (Any KD Mini-PROTEAN TGX Gel. BIO-RAD), followed by staining with SimplyBlue SafeStain (Invitrogene); the CPAP identity of the observed protein bands was confirmed by Western Blot (against the His-tag) and Mass spectrometry (MALDI TOF/TOF MS) analysis. The IMAC purified fractions were pooled to be used in further chromatographic steps.

3.1.1.4. Size exclusion chromatography (SEC) analysis

The IMAC purified protein was loaded into either preparative or analytical Superdex 200 columns equilibrated with cold buffer Hepes pH7.4, 300mM NaCl, 0.25mM TCEP and 10% Glycerol. The collected fractions were analyzed by SDS-PAGE (Any KD Mini-PROTEAN TGX Gel. BIO-RAD) stained with SimplyBlue SafeStain (Invitrogene). Some of the fractions were analyzed by negative staining transmission electron microscopy (NS-EM)

3.1.1.5. Ion exchange chromatography

Part of the IMAC purified protein was pooled and dialyzed for 3h at 4°C against buffer Hepes pH7.4, 150mM NaCl, 0.25mM TCEP and 10% Glycerol, and then applied on an anion exchange HiTrap Q HP column (GE Healthcare Life Sciences). The collected fractions were analyzed by NS-EM

3.1.2. MCM4/6/7

Mouse MCM4/6/7 protein sample was kindly provided by Dr. Hisao Masai, from the Department of Genome Medicine, Tokyo Metropolitan Institute of Medical Science, 2-1-6 Kamikitazawa, Setagaya-ku, Tokyo, 156-8506, Japan. The genes of MCM4-6xHis + MCM6 and MCM7-FLAG were coexpressed in a baculovirus system. MCM4/6/7 complexes were produced in High 5 insect cells coinfected with recombinant baculoviruses, and purified by successive steps of metal affinity chromatography, anti-FLAG M2 antibody-agarose affinity chromatography and, finally, glycerol gradient sedimentation, as described in (You & Masai 2008). The sample buffer was 50mM Hepes-Na, pH 7.5, 50mM Na-Acetate, 1mM DTT, 0.5mM EDTA, 2mM Mg-Acetate, 1mM ATP, 0.01% Triton X-100 and ~1.5% Glycerol.

3.1.3. Twinkle

Human mitochondrial replicative DNA helicase protein (Twinkle) samples were kindly provided by Prof. Laurie S. Kaguni, Distinguished Professor from Michigan State University, U.S.A and the Tampere University, Finland. The human full-length (residues 43-684) Twinkle-6xHis protein and an N-terminal deletion variant of Twinkle (6xHis-Del.E147K684) lacking the first 146 residues, where it is included the ZBD (ΔZBD) (Figure 15), were coexpressed in a baculovirus system. Both samples were produced in *Spodoptera. frugiperda* cells coinfected with recombinant baculoviruses, and purified by successive steps of metal affinity chromatography, heparin-sepharose affinity chromatography and ultracentrifugation in 12–30% glycerol gradients, as described in (Ziebarth et al. 2007). The purified samples were chromatographed on a Superdex 200 HR 10/30 column and analyzed by SDS-PAGE and Coomassie staining. Fractions corresponding with a mix of both hexameric and heptameric oligomers, as determined by size exclusion chromatography (SEC) analy*sis*, where used for our EM analysis. The buffer composition of the protein samples was 35 mM Tris-HCl, pH 7.5, 2 mM β-mercaptoethanol, ~5-20% glycerol, with 330 mM NaCl or 250 mM NaCl for the full-length protein and the ΔZBD, respectively. For some of the experiments conducted along this work, we decrease the salt concentration in a range between 100-150 mM NaCl, in order to allow for the binding of cofactors (4 mM $MgCl_2$ and 2 mM ATPγ-S).

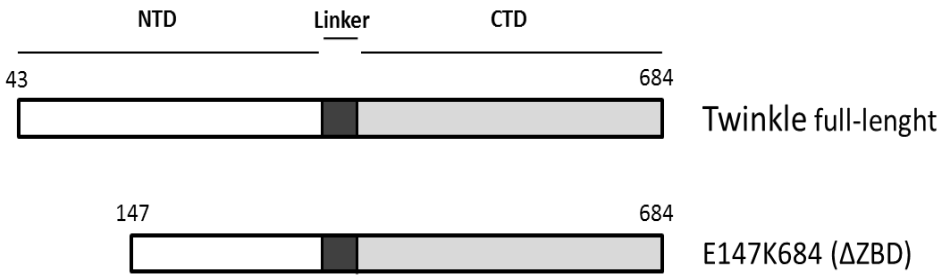


Figure 15.Schematic representation of Twinkle protein constructs. Mature full-length TWINKLE protein and Del.E147K684, a deletion construct lacking the ZBD (ΔZBD), where it is shown the position of the NTD (white), the NTD-CTD linker (Linker, in black) and the CTD (grey).

3.1.3.1. Solubility assay

Decreasing salt concentration: An initial sample reaction mixture containing 35 mM Tris-HCl, pH 7.5, 2 mM β-mercaptoethanol, 5% glycerol and 330 mM NaCl, was analyzed by negative staining electron microscopy (NS-EM). Separately, two other aliquots of the same reaction were directly diluted until reaching either 150 mM NaCl or 100 mM NaCl concentrations, while maintaining the same concentration of the rest of reagents and also adding 4 mM $MgCl_2$ and 2 mM ATPγ-S in both cases. The reaction samples were incubated for 10 min at 10 °C and subsequently analyzed by NS-EM.

Increasing salt concentration: An initial sample reaction mixture containing 35 mM Tris-HCl, pH 7.5, 2 mM β-mercaptoethanol, 5% glycerol, 4 mM $MgCl_2$, 2 mM ATPγ-S and 100 mM NaCl, was analyzed by NS-EM. Separately, the salt concentration of another aliquot of the same reaction was increased until 330 mM NaCl. The reaction was incubated for 25 min at 25 °C and subsequently analyzed by NS-EM.

3.2. In silico structural analysis: Sequence alignment and atomic structural modeling

3.2.1. CPAP$^{897\text{-}1338}$

Predictions of order/globularity and disorder regions of *Hs* CPAP (full-lenght) were done using the GlobProt2 server (Linding 2003). For CPAP$^{897\text{-}1338}$, the secondary structure was predicted with Jpred3 (Cole et al. 2008); for the prediction of coiled-coil regions the web server COILS (Lupas et al. 1991) was used; and, finally, the web servers PrOCoils (Mahrenholz et al. 2011) and LOGICOIL (Vincent et al. 2013) were employed to predict the oligomerization state of the coiled-coil part in the protein sequence. The atomic modeling of the CPAP coiled-coil CC4/CC5, as well as the CCGb-linker region, was carried out with I-TASSER, a protein structure/function prediction server that incorporates multiple-threading alignments, *ab initio* modeling and refinement by iterative assembly simulation of template fragments (Zhang 2008; Roy et al. 2010). The crystallographic structure of the G-Box/TCP domain (PDB-4BXP), together with the structures modeled for the CC4/CC5 and CCGb-linker regions, were used for both, proposing a model of a CPAP$^{897\text{-}1338}$ monomer and making a tentative atomic fitting in the EM reconstructed volume of a CPAP$^{897\text{-}1338}$ oligomeric complex.

3.2.2. MCM4/6/7

The sequences of a number of MCM homologs from different organism from both eukaryote and archaea, were compared with the mouse proteins by making multiple sequence alignments with the program ClustalX2 (Larkin et al. 2007). Atomic models for each of the mouse proteins MCM4, MCM6 and MCM7 were generated using the structural prediction server I-TASSER (Roy et al. 2010; Zhang 2008).

3.3. Transmission electron microscopy

Small, non-symmetrical flexible particles are difficult specimens for structural studies, although this heterogeneity may reflect the natural state required for their function, and so, study of these "hard" specimens is of high biological relevance.

3.3.1. Sample preparation and image acquisition

For negative staining (NS) of the samples, a drop of the protein solution was applied directly onto a glow-discharged EM grid (QUANTIFOIL. Formvar/Carbon. Cu 400 mesh grids), and allowed to be adsorbed on the grid surface (for few seconds or minutes, depending on protein concentration); then the drop was blotted with filter paper (Whatman grade No. 1) and the grid was washed by touching the surface with two consecutive drops of 0.75% (w/v) uranyl formate, blotting each time, and stained for 1min with one more drop of the same staining agent. Finally, the grid was blotted again and allowed to air dry before observation. Grids were examined in a JEOL JEM-1230 (accelerating voltage 100kV) electron microscope, and images recorded with a CCD camera ORIUS SC100 (4k x 2.7k pixel) at 40000x magnification, resulting in an image pixel size of 2.28 Å/pixel.

3.3.2. Image preprocessing

All image preprocessing, particle selection and two-dimensional (2D) analysis steps were carried out following the general workflow of the image processing package Xmipp3.1 (De la Rosa-Trevín et al. 2013). EM single particles were either manually or semi automatically selected.

Phase contrast is the dominant image formation mechanism for both stained and unstained specimens, and arises from the phase shift between scattered and unscattered electron waves (Amos et al. 1982). During EM data acquisition, the ideal projection image is convoluted with a point spread function (PSF) within the objective lens system. The PSF is a direct consequence of the lens aberrations, such as spherical (Cs) and chromatic aberrations (Cc) and, astigmatism (as result of imperfect construction of the lens), among others, plus certain degree of defocus combined with a small percenrage of amplitude contrast. The Fourier Transform of the PSF is the CTF (Contrast Transfer Function), which has a very characteristic pattern of cylindrical rings known as Thon-rings. A more in-depth discussion of the CTF and its analytical expression can be found in (Zhu & Frank 1997) and (Amos et al. 1982). Naturally, the blurring effect of the CTF must be

corrected, at least partially, before doing any further processing of the images. CTF estimation and correction was done with Xmipp3.1. CTF corrected images were downsampled by a factor of 2, reaching a final sampling rate of 4.56 Å/pixel. The selected particles were extracted and the images were normalized for its further processing.

Since electron micrograph i of biological particles are very noisy, the signal-to-noise ratio must be improved by performing the classification and averaging of large numbers of similar individual projections. Here, 2D class averages of particles were obtained with a clustering two-dimensional classification (CL2D) alignment and classification method (Sorzano et al. 2010), as implemented in Xmip3.1 (De la Rosa-Trevín et al. 2013). CL2D is a multi-reference refinement approach based on correntropy (a nonlinear similarity measure between two random variables), instead of the more traditionally used correlation, as similarity criterion. Particle classification results from the comparison of the set of correntropies of all the images assigned to one class, and the set of correntropies of all the images that do not belong to the same class. 2D classification was used for several purposes, such as: cleaning the image data set by eliminating particles of the worst 2D classes, dealing with heterogeneous population of particles and generating initial 3D reference models.

3.3.3. Density maps reconstruction: 3D classification and refinement

Single Particle Analysis of EM images have proved to be a powerful technique to elucidate the 3D structure of biological specimens within a wide range of different sizes and shapes, being the most appropriate choice when dealing with dynamic and heterogeneous samples. To obtain the 3D-EM map of a particle, an initial estimation/guess of its 3D structure is iteratively refined with a set of experimental 2D images (usually, several thousands). However, a phenomenon known as the "initial volume problem", consisting in the dependence of the final result on the initial input volume (promoted by a tendency of refinement algorithms to fall in local minimum) (Pascual-Montano et al. 2006), (Henderson et al. 2012), makes necessary to implement some kind of validation test. To handle the above mentioned problem, we used different initial volumes which were then refined with the same data set, so that only convergent maps, regardless of the initial volume, were finally selected. The initial volumes used included a blob of noise, an electron dense

sphere, as well as several volumes previously generated from our own experimental images with RANSAC (Vargas et al. 2014) or Random Conical Tilt (Radermacher 1988) (as implemented in Xmipp 3.1). These volumes were subjected to iterative refinements without applied symmetry and, in most cases, additional 2D and 3D classification runs were performed until obtaining convergent maps (Figure 16). The final validation for each of the obtained models was done by comparing projections of the 3D maps with experimental reference-free 2D class average classes.

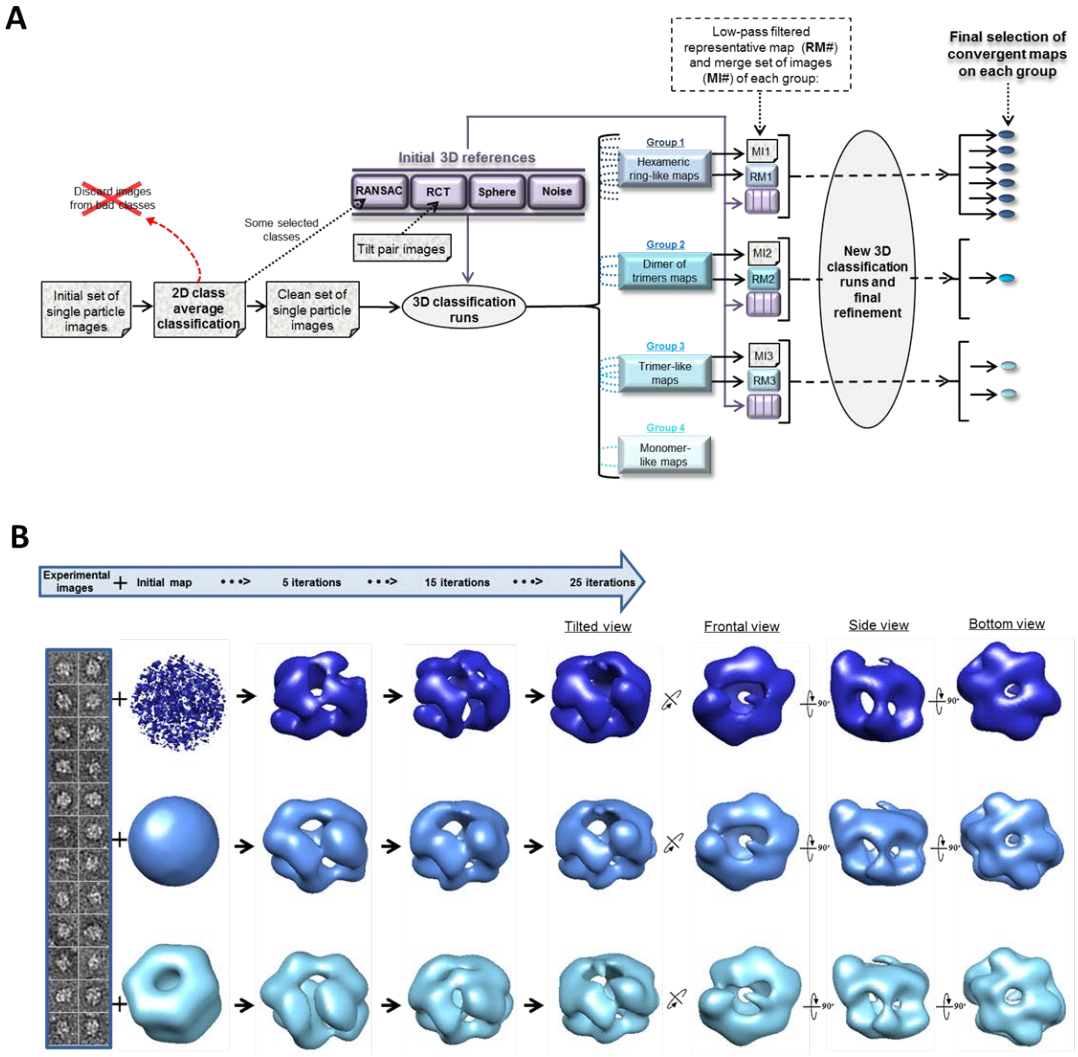


Figure 16. 3D reconstruction process. (A) Diagram of the general steps followed for 2D and 3D processing of EM images of single particles, using the data of the MCM4/6/7 subproject from the present work. **(B)** Example of an iterative volume refinement process and final map validation: A number of independent refinement runs are performed using the same data set of experimental images (*first* column), but starting from a different initial volume (*second* column).). The obtaining of a convergent structure validates the selected final map. In this example, the starting volumes are, from top to bottom: noise, a sphere and a150Å filtered model obtained with RANSAC (from experimental 2D average class images).

The resolution for each of the final 3D maps was determined using the Fourier shell correlation (FSC) 0.5 cutoff criterion (Heel & Schatz 2005). 3D maps where visualized with UCSF Chimera (Pettersen et al. 2004). The threshold level isosurface at which each density map was displayed to approximately represent the theoretical molecular mass, was calculated with the program EMAN1 (Ludtke et al. 1999) using the following command (http://blake.bcm.edu/emanwiki/Volume): volume <input file> <A/pix> [calc=<mass kDa>]. Here, the volume/mass conversions assume a protein density of 1.35 g/ml (0.81 Da/A3).

3.4. Normal Mode Analysis (NMA) of MCM4/6/7

Normal mode analysis (NMA) of EM structures has been extensively used in the study of flexible macromolecules and complexes for predicting their functional motions (Jin et al. 2014), (Tama et al. 2002), (Tama et al. 2003), (Ming et al. 2002), where putative dynamics is calculated directly from the structure. Using the NMA tool included in Xmipp3.1, six EM maps representing different conformational states of the ring-like MCM4/6/7x2 complex were analyzed. The distance matrix of the maps was calculated, for which it was employed the correlation among all the structures (similarity matrix) after doing the elastic alignment (deformations that one density map acquire to reach the shape of the next most similar 3D map) of the maps. A multidimentional scaling of the distance matrix was done in order to project all the structures into a 3D space, allowing a graphical visualization of the relative similarity distances among the six different maps of MCM4/6/7x2.

Chapter 4

4. Results

4.1. Human CPAP$^{897\text{-}1338}$

4.1.1. Sequence analysis and structural modeling

The *in silico* structural analysis of the *Hs* CPAP$^{897\text{-}1338}$ sequence (Figure 5A* and Figure 17) shows that there are three structurally different regions distributed as follows:

- 38% corresponds to the CC4/CC5 coiled-coil region and is predicted to be mostly formed by α-helices.
- 19% is formed by the CCGb-linker and is predicted to be unstructured.
- 43% forms the G-Box domain and is predicted to have 22 short β-sheets, which is in agreement with the equivalent part of the sequence solved for *D. rerio* in PDB-4BXP.

Figure 17. Secondary structural prediction of CPAP$^{897\text{-}1338}$.(A) The sequence includes the CC4/CC5 coiled-coil region (residues 897-1065), the CCGb-linker (residues 1066-1149) and the G-Box (residues 1150-1338). Red letter H referst to α-helices. Yellow letter E refers to β -sheets. **(B)** Sequence alignment of the G-Box domain from *H. sapiens* (*Hs*) and *D. rerio* (*Dr*), showing a comparison of the β-sheet positions for both the prediction for *Hs* (labelled in orange) and the already known structure from the PDB-4BXP for *Dr* (labelled in blue). The part corresponding to the *Dr* sequence underlined in red, is not resolved in the crystallographic structure.

Frequently, the interpretation of a 3D-EM density map involves building a molecular model. The structural modeling of the CC4/CC5 shows two ~12.5nm long α-helices organized in an antiparallel disposition, which is in agreement with the most probable oligomeric state predicted by the LOGICOIL algorithm. Although the CCGb-linker region is predicted to be intrinsically disordered, meaning that is flexible and there is not a unique/fixed structure, we still have used some of the structures modeled by I-TASSER as a guide to have a rough idea of the general dimensions for both potentially contracted or extended states of this zone of the protein. Finally, we have used the crystallographic structure of the G-Box/TCP domain from *D. rerio*, whose sequence presents ~68% identity with the G-Box from human (Figure 18).

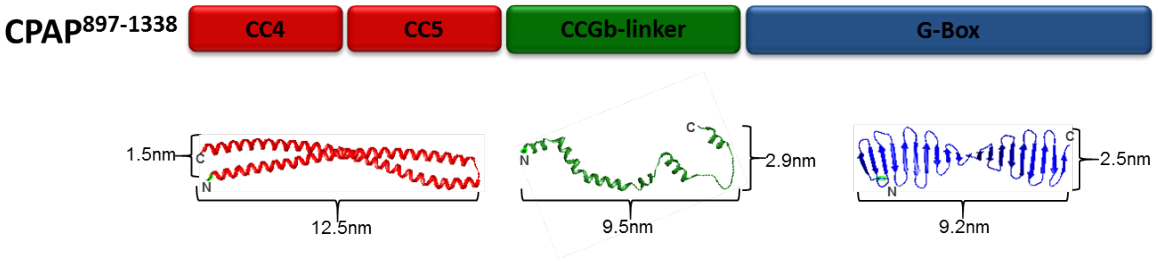


Figure 18. Color-coded schematic representation of *Hs* CPAP$^{897-1338}$ protein´s domains (*up*) and their related atomic model (down). From left to right: I-TASSER predicted atomic models for both the CPAP coiled-coils CC4/CC5 (in *red*) and an extended state version of the flexible CCGb-linker (in *green*), and the crystallographic structure of the G-Box/TCP domain (PDB-4BXP) (in *blue*).

4.1.2. Protein purification

Most parts of CPAP are predicted to be disordered and coiled-coil, which are two hallmark characteristics of many centrosomal proteins, and are commonly associated with samples highly insoluble and difficult to handle and purify (Dos Santos et al. 2013)(Treviño et al. 2014), which unfortunately is the case of the full-length CPAP. The CPAP$^{897-1338}$ sequence codes for a protein with a theoretical molecular weight (MW) around 52 kDa. This construct includes the only globular domain of CPAP (the G-Box), and the rest of the fragment, which covers around 57% of its sequence, is composed by the coiled-coils CC4/CC5 and the predicted non-structured region CCGb-linker (which connects CC4/CC5 with the G-Box).

Our N-terminal 6xHis-tagged CPAP$^{897-1338}$ construct is partially soluble under defined optimized expression and purification conditions (more details can be found in Materials and Methods). A one-step IMAC purification gave rise to a highly pure protein sample (Figure 19 A and B). SDS-PAGE analysis of the purified samples showed a small amount of a discrete degradation pattern of the protein; the CPAP identity of the observed lower molecular weight bands was confirmed by Mass spectrometry analysis and Western blot against the N-terminal His-tag of the protein, which revealed that the degradation occurs by the C-terminus side (Figure 19C).

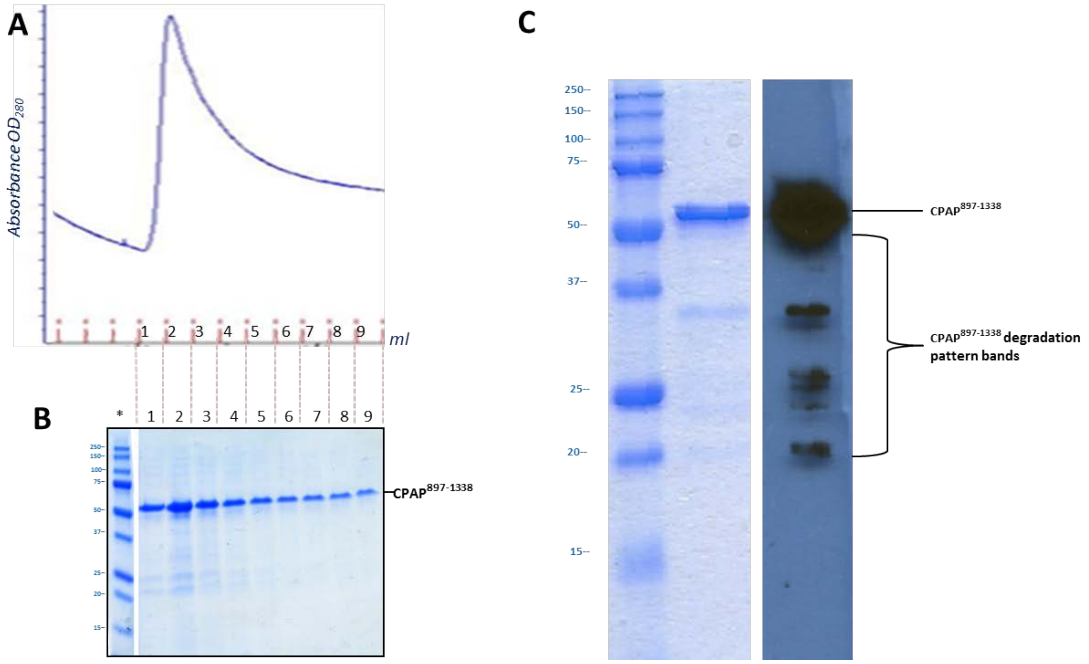


Figure 19. Immobilized metal affinity chromatography (IMAC) purification of CPAP$^{897\text{-}1338}$. (A) IMAC purification chromatographic elution profile and **(B)** SDS–PAGE of the purified fractions. The lane labelled with the asterisk (*) corresponds to the protein molecular weight markers; the lanes 1 to 9 are the pooled fractions of the purified CPAP$^{897\text{-}1338}$. Dashed lines show the fraction sample correspondence between the chromatogram and SDS-PAGE. **(C)** CPAP$^{897\text{-}1338}$ protein degradation pattern. SDS-PAGE (*left*) and Western Blot against the N-terminal His-CPAP$^{897\text{-}1338}$ (*right*) showing a characteristic but minoritarian protein degradation pattern observed along the purification.

The variable SEC profiles obtained in a number of purifications of CPAP$^{897\text{-}1338}$ performed under the same buffer and temperature conditions, indicated differences in size and abundance of the macromolecular specimens that were present. A possible explanation to this phenomenon is the change in concentration of the initial input samples passed through each of the SEC runs, which would strongly suggest that the concentration of the protein must be a determinant factor on the oligomerization state of CPAP$^{897\text{-}1338}$, thus causing variations on the abundance and hydrodynamic radius of the particles. The relatively broad elution profile obtained in some cases is explained by the coexistence of particles with different shape/size in the sample; this was confirmed by EM analysis of the eluted fractions.

4.1.3. Structural characterization of different CPAP$^{897\text{-}1338}$ homo-oligomeric complexes and fitting of atomic models

Visualization of NS-EM images of a number of CPAP897-1338 purified samples allowed us to identify different complexes formed by this protein. Indeed, we were able to characterize different complexes of varying size, corresponding to different oligomeric structures. It is important to note that in the following we will perform a "tentative/putative" assignment of oligomeric state based on both the biochemical data previously presented as well as low resolution atomic fittings. Note that our fittings are only intended to show how the different maps are perfectly compatible with the oligomeric state assigned to them, and not to depict precise atomic arrangements.

4.1.3.1. CPAP$^{897\text{-}1338}$ flexible, fibrillar structures

EM of a SEC (Superdex 200 16/60) purified CPAP$^{897\text{-}1338}$ protein fraction eluted around an apparent MW of 120kDa (Figure 20A), showed a population of flexible and elongated rope-like structures (Figure 21). Due to the fibrillar shape of these particles, it is expected that their hydrodynamic behavior significantly differ from that of the globular molecular weight protein markers commonly used for SEC columns calibration. A Native-PAGE was run to determine the oligomeric state of the protein in this fraction (Figure 20 B), showing that the predominant state was the monomer (the protein runs around 66kDa under non-denaturing conditions) and only a negligible part forms some higher order oligomers (CPAP identity of the bands was confirmed by Mass spectrometry analysis). Protein degradations products similar to those previously observed (Figure 19C) were also present in this fraction.

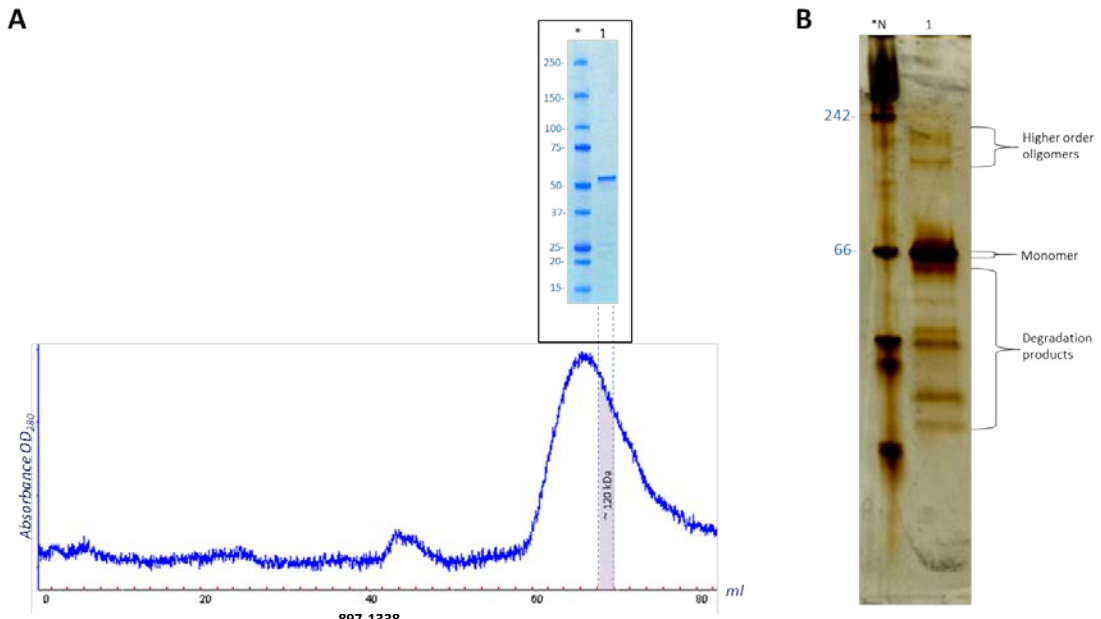


Figure 20. SEC purification of CPAP$^{897\text{-}1338}$ fibers. (A) Size-exclusion chromatography (SEC) profile (Superdex 200 16/60) of purified CPAP$^{897\text{-}1338}$. Black box shows an SDS–PAGE image, where the lane labeled with the asterisk (*) corresponds to the protein molecular weight markers and, the lane labeled with the number 1 represents a fraction that elutes at an apparent MW of ~120kDa. EM images of this fraction show flexible fiber particles (See Figure 21). **(B)** Native-PAGE (silver staining) of the same fraction showed in panel A (labeled with the number 1). Most of the protein is in a monomeric state (thick band close to the 66kDa marker), although there are some very soft bands corresponding to a number of oligomers (Over the 66kDa band) and some others bands showing the previously observed protein degradation products (under the 66kDa band). *N, shows the lane of the native protein ladder.

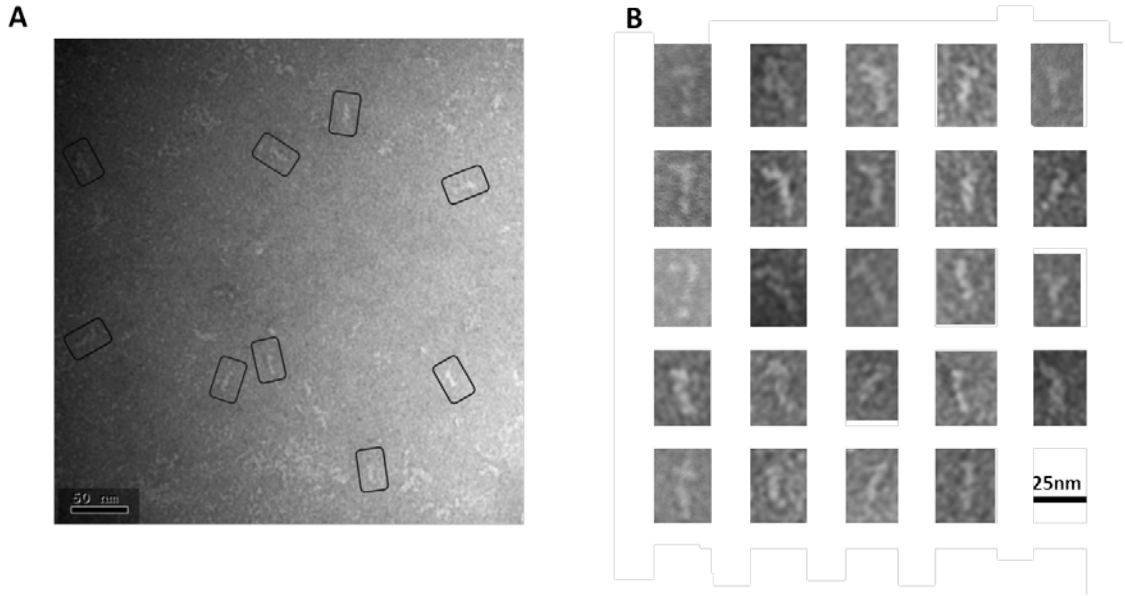


Figure 21. EM of CPAP$^{897\text{-}1338}$ fibers. (A) A representative negative stain electron micrograph of a SEC purified CPAP$^{897\text{-}1338}$ fraction eluted at an apparent MW of ~120kDa. It shows a large population of flexible fiber-like particles. Some selected particles are highlighted by black rectangles. **(B)** Windowed raw images of some side views show a general elongate cane-like structure.

Although the high flexibility of the particles made unfeasible to obtain well-defined 2D average classes of the EM images to make a 3D reconstruction, it was still possible to carry out a partial structural analysis by measuring the general dimensions of the most representative images, which presented a cane-like shape with a "shaft" of~15-17nm long and a flexible "handle" of~10nm long; the apparent thickness of the strings varied in a range between 2-4nm.

To complete the assembly of an atomic model for a possible 3D structure of the CPAP$^{897-1338}$ monomer, we put together the crystallographic structure of the G-Box/TCP domain (PDB-4BXP) and the models of the CC4/CC5 and the CCGb-linker (an extended version) regions, as predicted by I-TASSER; here, the CCGb-linker acts as a very flexible hinge that connects the G-Box and CC4/CC5 domains, which respectively correspond to the "shaft" and to the "handle" of the model (Figure 22 A). We observed that the dimensions and the shape of some of the fiber-like EM images fitted reasonably well with the model proposed before (Figure 22 B).

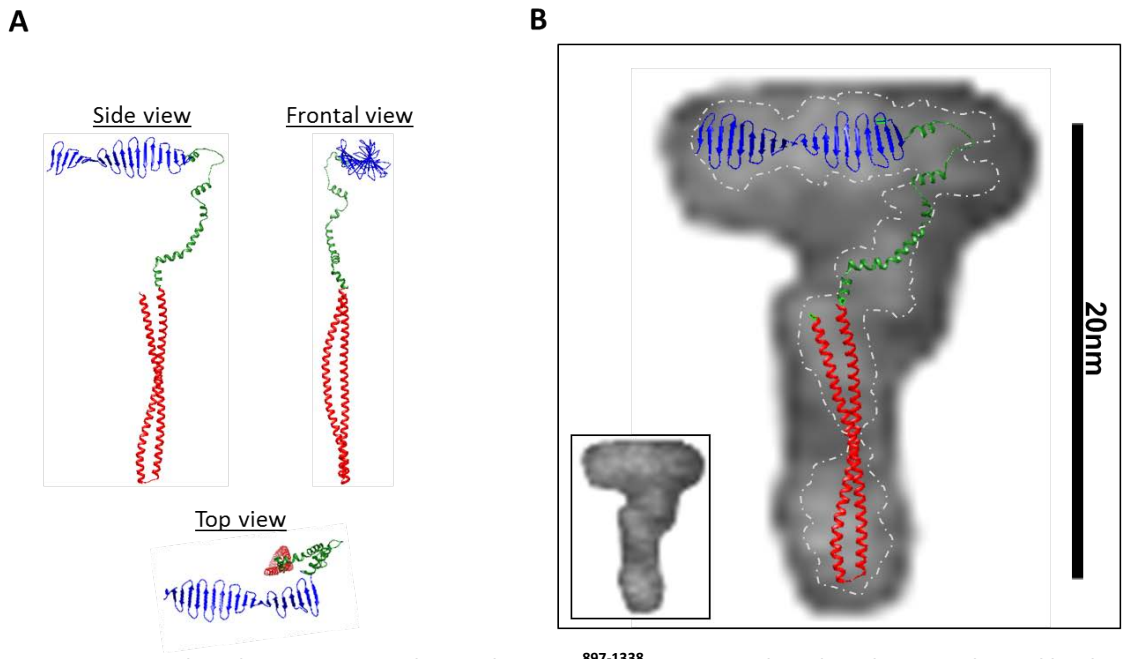


Figure 22. Structural prediction corresponding to the CPAP$^{897-1338}$ monomer, based on the general cane-like shape of EM images from a SEC purified sample. (A) Model of a possible CPAP$^{897-1338}$ monomer assembled using structural predictions of the CC4/CC5 and CCGb-linker domains and the X-ray structure of the G-Box (PDB-4BXP). **(B)** Superimposition of the atomic models of the CC4/CC5, CCGb-linker and G-Box domains on a representative single particle image from a NS-EM sample of CPAP$^{897-1338}$. A discontinuous line demarks the contour of the EM particle. A smaller version of the same particle is shown in the inset at the lower left corner.

4.1.3.2. CPAP$^{897\text{-}1338}$ toroidal complex

EM micrographs of a SEC (Superdex 200 10/300) purified sample eluted at an apparent MW of ~108kDa (Figure 23), which is compatible with the size of a CPAP$^{897\text{-}1338}$ dimer, was mostly populated by both images of toroids with an external diameter of ~10nm, and a more rectangular kind of particles with a size of ~7.5nm (*height*) x 10nm (*width*), which look like two wavy threads bound by their wider sides (Figure 24 A and B). These images could be interpreted as different projections of a structure similar to two interweaved rings, where the rectangular-like and the ring-like particles could be respectively attributed to the lateral and top views of the proposed model. Some of these particles showed thin flexible strings protruding to the outside (Figure 24 C) and, interestingly, we observed pairs of particles interacting through these flexible extensions (Figure 24 D).

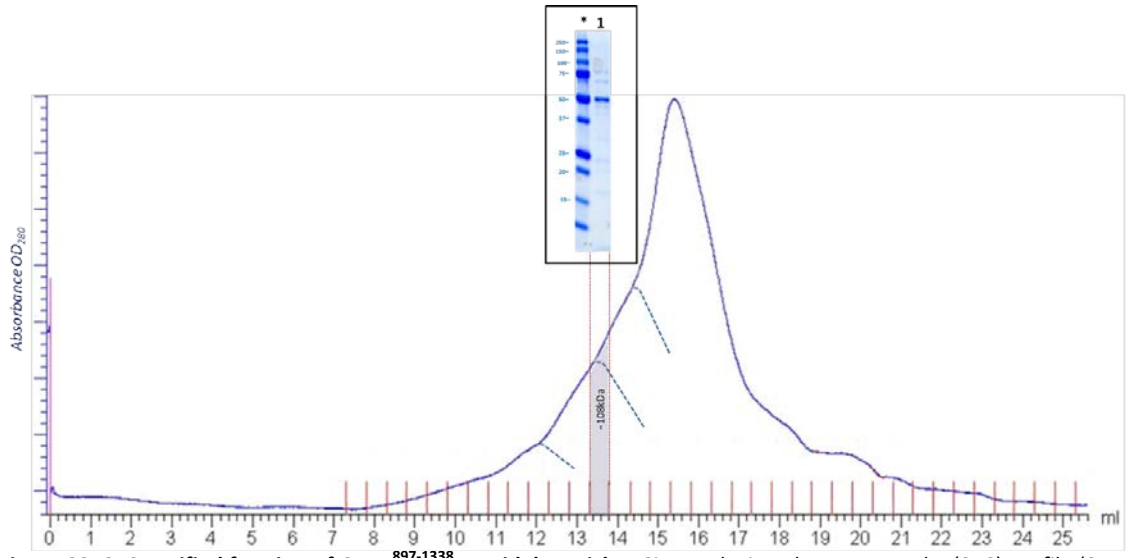


Figure 23. SEC purified fraction of CPAP$^{897\text{-}1338}$ toroidal particles. Size-exclusion chromatography (SEC) profile (Superdex 200 10/300) of purified CPAP$^{897\text{-}1338}$ where blue dotted lines mark three fused peaks under the graphic. The black box shows an SDS–PAGE image, where the lane labeled with the asterisk (*) corresponds to the protein molecular weight markers and the lane labeled with the number 1 is a fraction that elutes at an apparent MW of ~108kDa, compatible with a dimer of CPAP$^{897\text{-}1338}$. EM images of this fraction show some toroidal particles (See Figure 24).

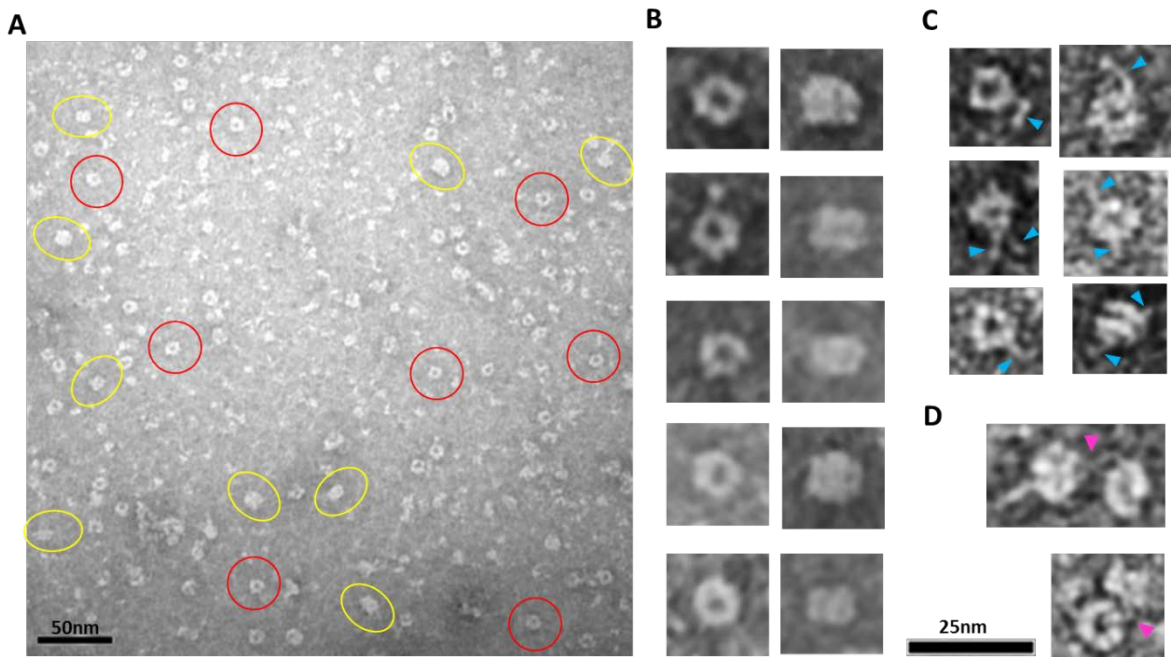


Figure 24. EM of a SEC fraction of CPAP$^{897\text{-}1338}$ eluted at an apparent MW around 108kDa. (A) A representative negative stain electron micrograph of a SEC purified CPAP$^{897\text{-}1338}$ fraction eluted at ~108kDa, compatible with being a dimer of the protein. Some toroid-like views are highlighted with red circles, while the more rectangular particles are highlighted by yellow ovals. **(B)** Windowed raw images of representative particles that could correspond with top views (*left panel*) and lateral views (*right panel*) of the same type of particle. **(C)** Single particles showing extended flexible projections (pointed by blue arrowheads). **(D)** Pair of particles interacting through flexile strings (pointed with a fuchsia arrowhead).

It is expected that during the averaging of the images, highly flexible structures get reduced, at best, to the base part, for being it the most stable point. Indeed, in agreement with the aforementioned phenomena, the 2D class average classification of 36,699 single particle images (Figure 25 A) showed some toroidal particles presenting a punctual, strong extra density at the periphery of the ring (Figure 25 B), which could be attributed to the average of the base part of the flexible strings observed in the single particle images (Figure 24 C), although it cannot be excluded that more compact conformations of the flexible projection exist. A small discontinuity is observed in some of the ring-like images, suggesting the presence of a gap (Figure 25 B). 2D image classification revealed the presence of other more elongated and flat particles (Figure 25 C), some of which looked like two small consecutive rings or an infinity symbol (∞), suggesting that other intermediate conformations of the dimer could also exist.

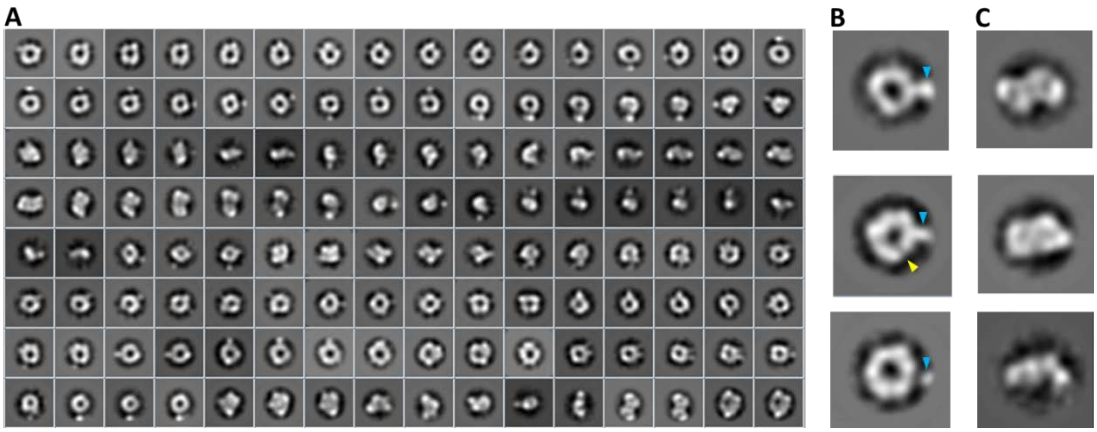


Figure 25. 2D image analysis of a sample fraction eluted at an apparent MW compatible with a CPAP$^{897-1338}$ dimmer. **(A)** Reference-free class average classification (CL2D) of 36,699 particle images. **(B)** Toroidal class images showing an extra mass at the periphery (pointed with a blue arrowhead), which is not always clearly connected to the ring. A weak density area (pointed with a yellow arrowhead) suggests a gap in the ring. **(C)** 2D class images of more elongated and flat particles that could correspond to intermediate conformation of the final dimer.

The combination of both 2D and 3D classification image processing procedures allowed to obtained a final 24 Å resolution map of an asymmetrical toroid-like structure (Figure 26) representing the most frequent kind of particles (Figure 24 B) of the putative dimer of CPAP$^{897-1338}$. The top and bottom openings of the central channel have different shape and size. The frontal view of the complex shows a slightly concave cross brace structure where two copies of the crystallographic structure of the G-Box/TCP domain (PDB-4BXP) can be reasonably fitted (Figure 26 B). The opposite side (back view) also has an X-like shape but is thicker and convex. Finally, each of the two remaining opposite sides shows a kind of two-tier like shape with an elongated opening in the middle. When the 3D map is displayed at a lower contour level (Figure 26 C) it appears an additional density emerging from a corner of the frontal face of the structure. This extra mass may correspond to part of one of the flexible strings previously described in some of the single particle images (Figure 24 C).

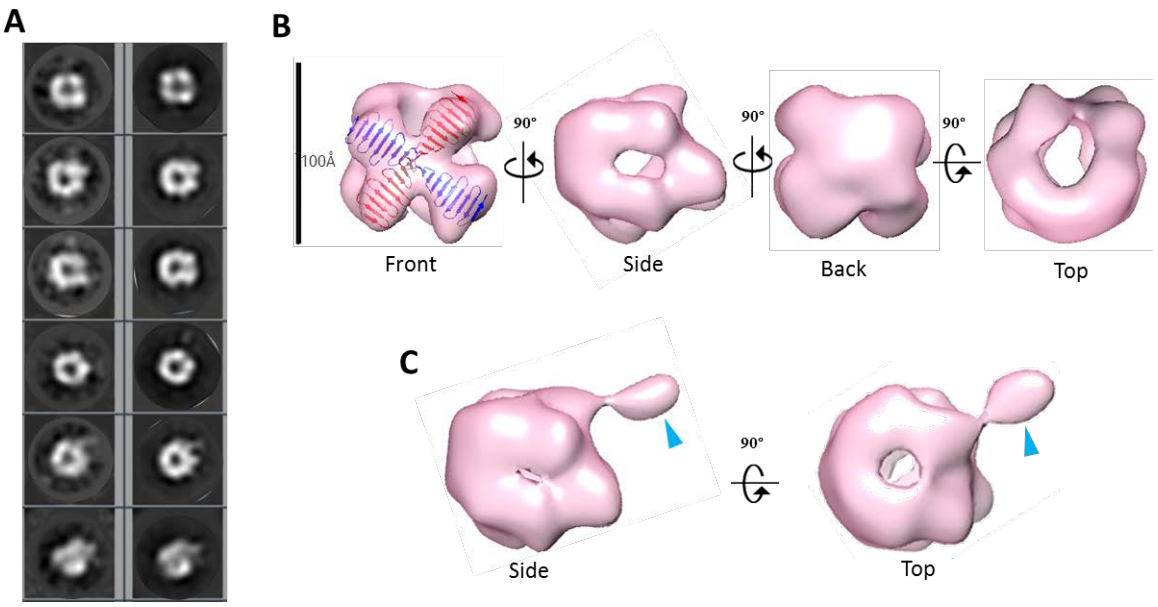


Figure 26. 3D reconstruction of a putative CPAP$^{897\text{-}1338}$ homo-dimeric complex. (A) Reference-free class averages (*left column*) and the corresponding forward projections (*right column*) of the 3D reconstruction. **(B)** Different views of the putative CPAP$^{897\text{-}1338}$ dimer with two copies of the atomic coordinates of G-Box/TCP domain (PDB-4BXP, represented in blue and red) fitted inside. **(C)** Top and one side view of the density map displayed at a low contour level where it is observed an extra mass (pointed by a blue arrowhead).

4.1.3.3. CPAP$^{897\text{-}1338}$ barrel-meshed like structure

A gel Filtration (Superdex 200 10/300) purified protein fraction eluted at an apparent MW of ~208kDa (Figure 27), compatible with the size of a tetramer of CPAP$^{897\text{-}1338}$, was analyzed by NS-EM (Figure 28 A). Reference free average classification of 3,048 singles images showed a meshed structure around 14.5nm length, that presents two different types of side views, each around 9.3nm and 8.3nm maximum width. The wider side view has an ellipsoid shape, with two transverse and slightly curved bridges (Figure 28 B). The narrowest side view looks like a lattice formed by strings with a diameter around 2.5nm, which agrees with the diameter of the crystallographic structure of the G-Box/TCP domain (PDB-4BXP). We also noted that the length of the structure of the G-Box (around 9.2nm) matches with the longitude of the diagonal lines observed in some of the 2D average class (Figure 28 C).

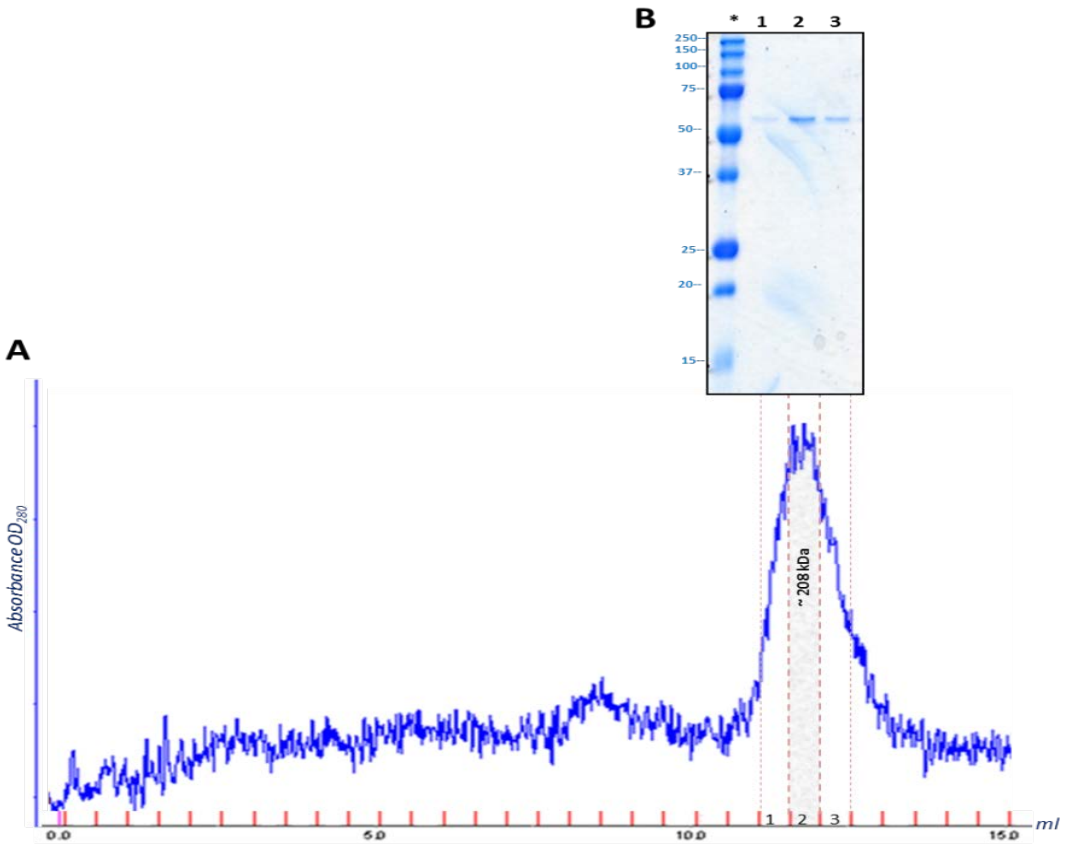


Figure 27. Purification of CPAP$^{897\text{-}1338}$ barrel-like particles. Chromatogram and associated SDS-PAGE are shown. **(A)** SEC profile (Superdex 200 10/300) of a purified CPAP$^{897\text{-}1338}$ sample. The center of the peak (fraction 2) elutes at an apparent MW of ~208kDa, compatible with a tetramer of CPAP$^{897\text{-}1338}$. This fraction was analyzed by EM (See Figure 28). **(B)** SDS–PAGE. Lane labelled with the asterisk (*) corresponds to the protein molecular weight markers; lanes 1-3, fractions corresponding to the pick of the purified CPAP$^{897\text{-}1338}$.

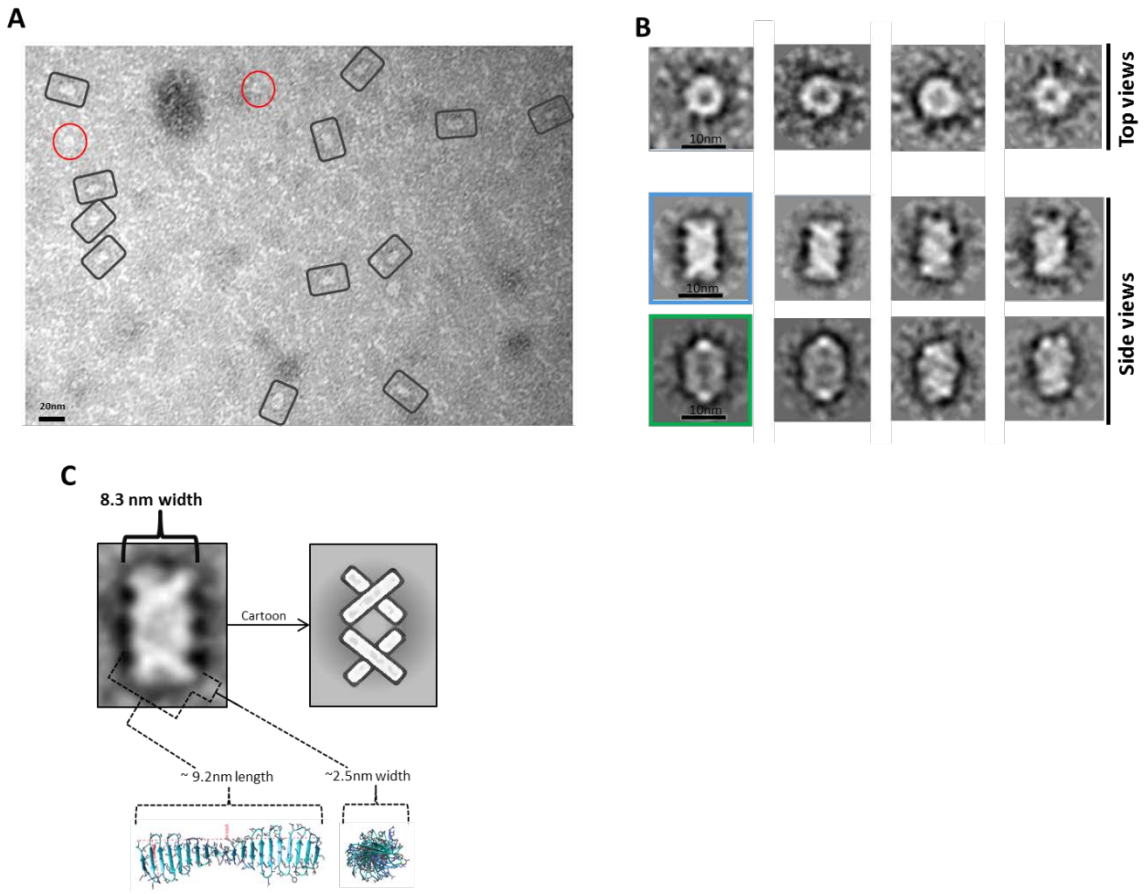


Figure 28. Transmission electron microscopy of negatively stained CPAP$^{897\text{-}1338}$ barrel-like particles. (A) A representative negative stain electron micrograph of a SEC purified CPAP$^{897\text{-}1338}$ fraction eluted at ~208kDa, which is compatible with being a tetramer of the protein. Top views and side views are highlighted by gray rectangles and red circles, respectively. **(B)** Reference-free class averages of top views (panel of the first row) and side views (panels of the last two rows) of the putative CPAP$^{897\text{-}1338}$ tetramer complex. Some of the side view images are almost rectangular and narrower (image highlighted in blue) than other kind of images that have a more oval-shape and are wider (image highlighted in green). **(C)** 2D class average of a rectangular side view of 8.3nm width, which resembles two stacked crossbars (see the schematic representation to the right of the image), where the match between the dimensions of its diagonal strings (~2.5nm x ~ 9.2nm) and the ones of the crystallographic structure of *D. rerio* G-Box/TCP domain (PDB-4BXP) is shown.

The 24 Å resolution 3D map of the putative CPAP$^{897\text{-}1338}$ tetramer reveals the structure of an asymmetrical barrel-like complex made of intertwined strings. A tentative fitting of the atomic model of the G-Box (PDB-4BXP) shows how the general dimensions of this crystallographic structure fits in our EM volume (Figure 29).

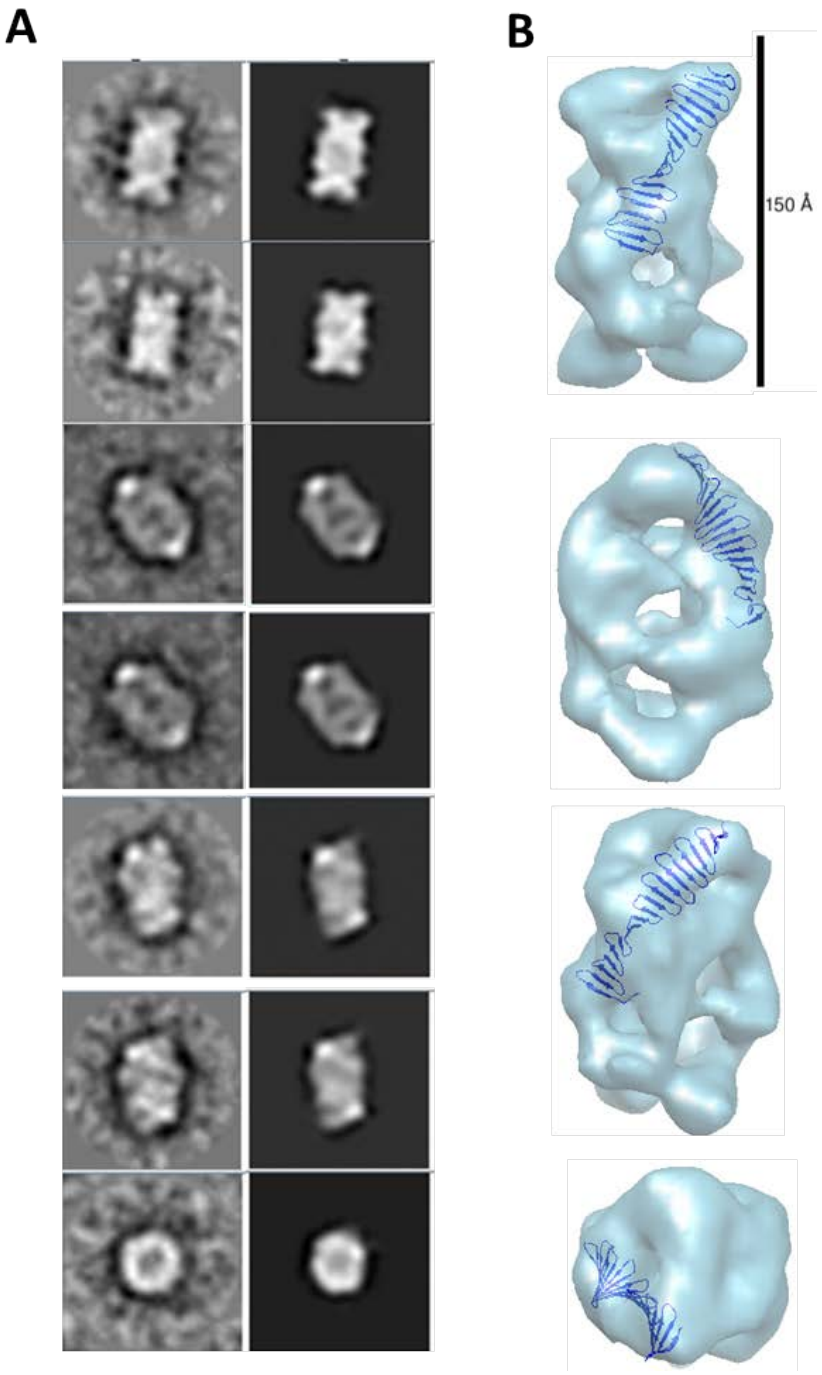


Figure 29. 3D reconstruction of a putative CPAP$^{897\text{-}1338}$ homo-tretrameric complex. (A) Reference-free class averages (*left column*) and the corresponding forward projections (*right column*) of the 3D reconstruction. **(B)** Different orientations of the putative CPAP$^{897\text{-}1338}$ tetramer volume, together with a tentative fitting of the atomic structure of the G-Box/TCP domain showing how the general dimensions of this structure fit into the EM volume.

Unstructured regions, as it is predicted to be the CCGb-linker of CPAP (Figure 17 A), use to be highly flexible, which allows the protein to acquire different conformations. An alternatively proposed conformation of the CPAP897-1338 atomic model presented in Figure 22, where the CC4/CC5 domain bends toward the G-Box producing a triangular-shape structural arrangement (Figure 30 A), allowed a nice fitting of four copies of the monomer inside the volume of our putative CPAP897-1338 tetramer (Figure 30 B and C). Step by step, the fitting has been done by distributing the structural domains of the protein monomers into the four long lateral faces of the density map, in the following way:

- G-face: Four copies of the crystallographic structure of the G-Box domain have been fitted in one of the narrow lateral side views of our map.
- CC-face: Four copies of the atomic model of CC4/CC5 have been fitted on the opposite side.
- Linker-face1 and linker-face2: Two copies of an extended version model for the CCGb-linker were accommodated at each of the two wider lateral sides of the volume.

The fitting described above gives additional support to the suggested tetramer oligomeric state for this complex. The resulting distribution of all the domains within the density map creates a clear structural polarity.

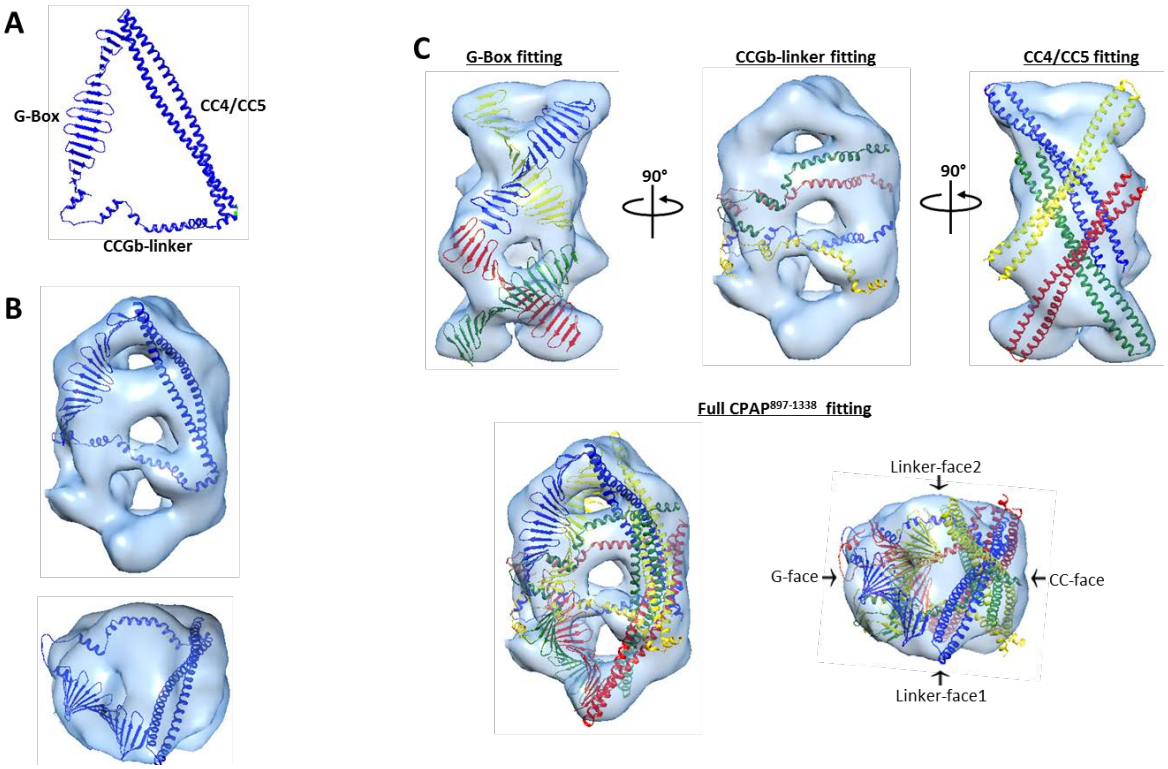


Figure 30. Fitting of four copies of a triangular conformation of the proposed atomic structure of CPAP$^{897\text{-}1338}$ into the 3D map of the putative CPAP$^{897\text{-}1338}$ tetramer. (A) Proposed triangular conformation of CPAP$^{897\text{-}1338}$ monomer, using the obtained atomic models of the CC4/CC5 and CCGb-linker domains and the X-ray crystallographic structure of the G-Box (PDB-4BXP). **(B)** Fitting of one copy of the complete monomer model into the 3D volume, where two different views of the EM density map of the complex are shown in translucent blue. **(C)** For clarity, in these figures the three structural domains of each of four fitted CPAP897-1338 monomers are painted with a different color. In the first row there are shown different views of the density map with the fitting of four copies of each of the structural domains, and in the second row it is shown the fitting of four copies of the complete monomer model into the 3D map. Abbreviations used for labeling different faces of the figure in the panel at the bottom right corner are: G, G-Box; Linker, CCGb-linker and CC, CC4/CC5.

Interestingly, the obtained 3D map of particles in the fraction eluted at a MW ~108 kDa, (which we proposed previously to be a dimer), presents both dimensions and shape that could be compatible with half of our putative tetramer 3D map. The above mentioned consideration, together with the observed images of two molecules of the putative dimers interacting with each other (Figure 24C), suggest that the tetramers may be formed by the dimerization of the ~108kDa toroidal complexes.

4.1.3.4. Modular higher order supramolecular rope-like structures

An aliquot of the IMAC purified CPAP$^{897\text{-}1338}$ sample was dialyzed to reduce the buffer salt concentration from 300 mM NaCl to 150 mM NaCl; the protein was subsequently passed through an anion exchange column. In several of the eluted fractions we observed thick rope-like structures of different lengths. At first glance, these structures seem to be formed by linear stacks of a variable number of rectangular modules with dimensions varying between ~7-8.5 nm (*length*) x ~15-16 nm (*width*) (Figure 31A and B), whose shape and size are similar to the corresponding ones of the narrow side views of our previously described putative CPAP$^{897\text{-}1338}$ tetramer. Interestingly, we observed that the dimensions of completely or partially isolated blocks (Figure 31C), are also very similar to the ones of the putative tetrameric complex previously described. Then, a possible explanation for the slightly variable internal organization of the modular rope-like structures is that they were formed by stacking of putative tetramers, allowing for some degree of flexibility of these blocks which could explain the small differences in length and width among the modules.

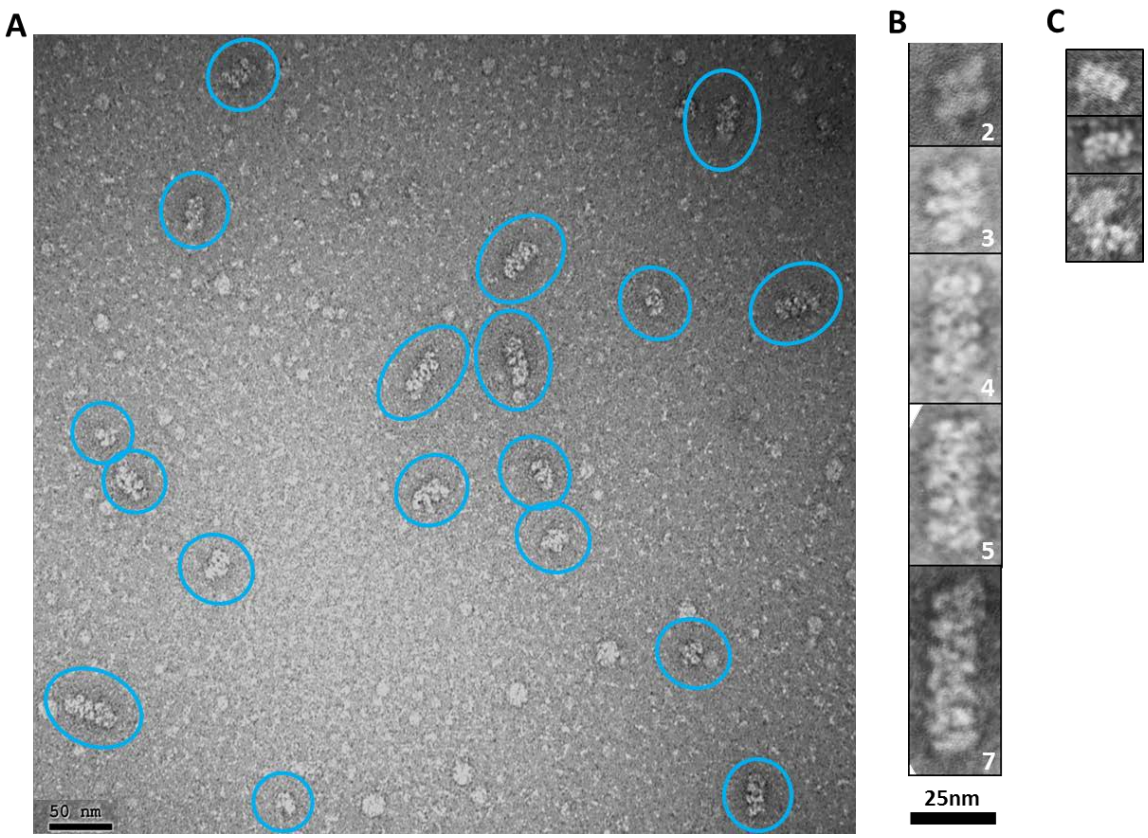


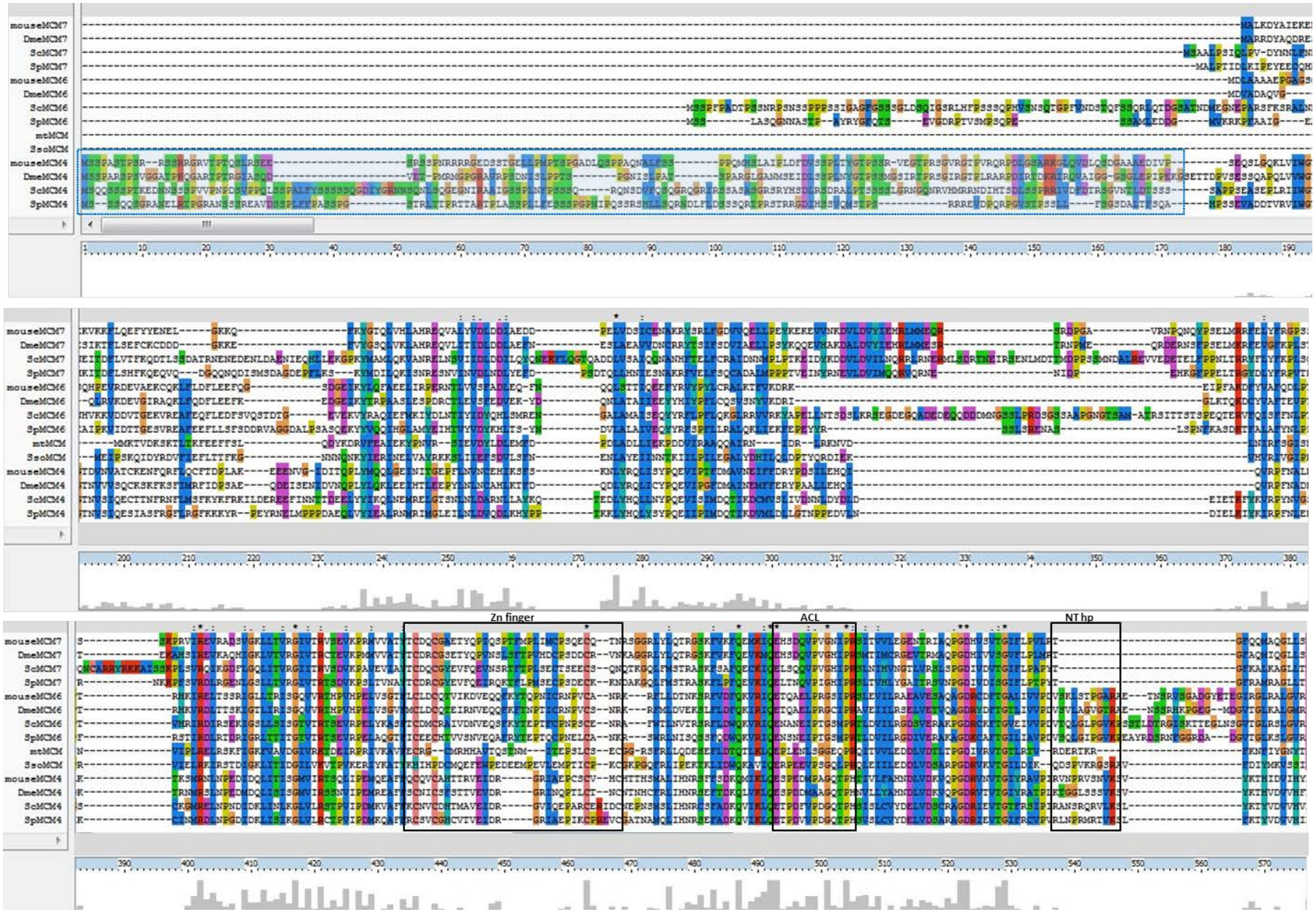
Figure 31. CPAP$^{897\text{-}1338}$ putative tetramers may stack together by their longer sides to form modular higher order complexes of variable length. (A) Representative negative stain electron micrograph of an anion exchange purified CPAP$^{897\text{-}1338}$ fraction, where there are observed higher order modular-like complexes of different length (highlighted by blue ovals). **(B)** Windowed zoomed images of some selected particles where the number of stacked modules is indicated by the number written on the lower right corner of each image. **(C)** Images of individual modules (first two rows) or partially interacting modules (last row).

It has been proposed that higher order assembles of CPAP could act in a synergistic way as a platform to provide an interface mediating the tethering of PCM to the centriole (Gopalakrishnan et al. 2011; Hatzopoulos et al. 2013; X. Zheng et al. 2014). Our 3D-EM reconstruction of the CPAP$^{897\text{-}1338}$ tetramer and the observed higher order extended and modular organized structures with an ~8 nm axial periodicity, are compatible with the last mentioned idea.

4.2. Mouse MCM4/6/7

4.2.1. Multiple sequence alignment and structure prediction modeling of mouse proteins MCM4, MCM6 and MCM7

To date, multiple biochemical and structural studies on MCM proteins from different organisms are available. We performed a multiple sequence alignment of mouse MCM4, MCM6 and MCM7 sequences against those from other species (Figure 32). The results of this analysis are highly informative and useful to get a better picture of the level of structural and functional similarity that can be expected between mouse proteins and the available information from other MCM homologs. It is observed that most of the structural motifs within the AAA^+ domain are well conserved, while the N-terminal and C-terminal regions are quite variable. When compared against most of MCM7 sequences (and with the *Sso*MCM), it become noticeable that MCM6 proteins present a C-terminal extension, as it happens with some of the MCM4 sequences, although to a much lesser extent. Furthermore, the MCM4 proteins display a long extension of residues toward the N-terminus.


mouseMCM7
DmeMCM7
ScMCM7
SpMCM7
mouseMCM6
DmeMCM6
ScMCM6
SpMCM6
mtMCM
SsoMCM
mouseMCM4
DmeMCM4
ScMCM4
SpMCM4
Zn finger
ACL
NT hp

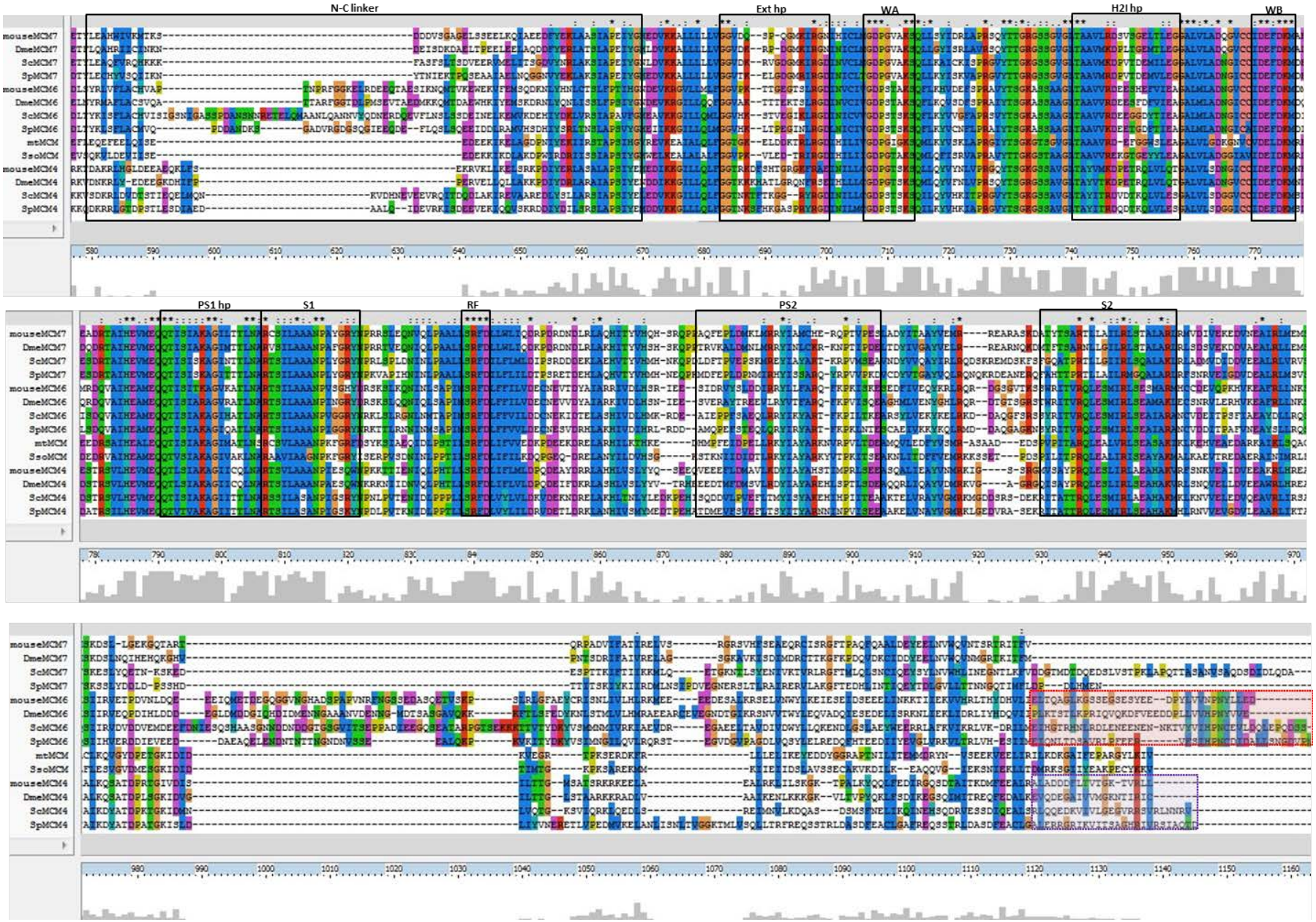

N-C linker
Ext hp
WA
H2I hp
WB
PS1 hp
S1
RF
PS2
S2
mouseMCM7
DmeMCM7
ScMCM7
SpMCM7
mouseMCM6
DmeMCM6
ScMCM6
SpMCM6
mtMCM
SsoMCM
mouseMCM4
DmeMCM4
ScMCM4
SpMCM4

Figure 32. Multiple sequence alignment of mouse MCM4, MCM6 and MCM7 complete sequences with respect to their homologs in representative organisms. *D. melanogaster* (*Dme*), *S. cerevisiae* (*Sc*), *S. pombe* (*Sp*), and the MCM protein in *S. solfataricus* (*Sso*) and in *M. thermautotrophicus* (*mt*). Shared motifs among MCM proteins are highlighted into black boxes labeled with the respective motif abbreviation: Zn finger, Zinc finger; ACL, allosteric communication loop; NT hp, N-terminal β-hairpin; N-C linker, amino and carboxyl domains linker; Ext hp, external β-hairpin; WA, Walker A motif; H2I hp, helix 2 insert β-hairpin; WB, Walker B motif; PS1 hp, presensor 1 β-hairpin; S1, sensor 1; RF, arginine finger motif;. PS2, presensor 2 insertion; S2, sensor 2. Residues conservation is highlighted by ClustalX color code and the light gray columns under the alignment are an indicator of the conservation level at each point (Chenna 2003),(Larkin et al. 2007).Blue dashed box highlights a N-terminal MCM4 residue extension zone; red dashed box highlights a C-terminal MCM6 residue extension zone; purple dashed box highlights a C-terminal MCM4 residue extension zone.

Consistently with the conserved domains and variable regions observed in the multiple sequence alignment of MCM homologs, the I-TASSER modeled structures of mouse MCM4, MCM6 and MCM7 maintain a core globular architecture similar to that of the crystal model of the near full-length *Sso*MCM (PDB-3F9V), which is lacking its C-terminal domain. On the other hand, length differences between proteins are observed at both their N-terminal and C-terminal parts. Mouse MCM6 has a long C-terminal extension and shares 95% of sequence identity with the structure of the C-terminus of *hs*MCM6 (PDB-3F9V), which is a winged helix (WH) motif. MCM4 features both a very long and unstructured N-terminal extension and a flexible but much shorter C-terminus, which is similar in length to that of archaea MCM. Sequence identity between the C-terminal region of mouse MCM4 and the solved C-terminals structures of *Sso*MCM (PDB-2M45) and *Mth*MCM (PDB-2MA3) are around 16% and 20%, respectively. Compared with the other proteins, the C-terminus of MCM7 is the shortest one; the structural prediction modeling of this part adopts a HTH conformation that folds down against the AAA^+ domain of the protein (Figure 33).

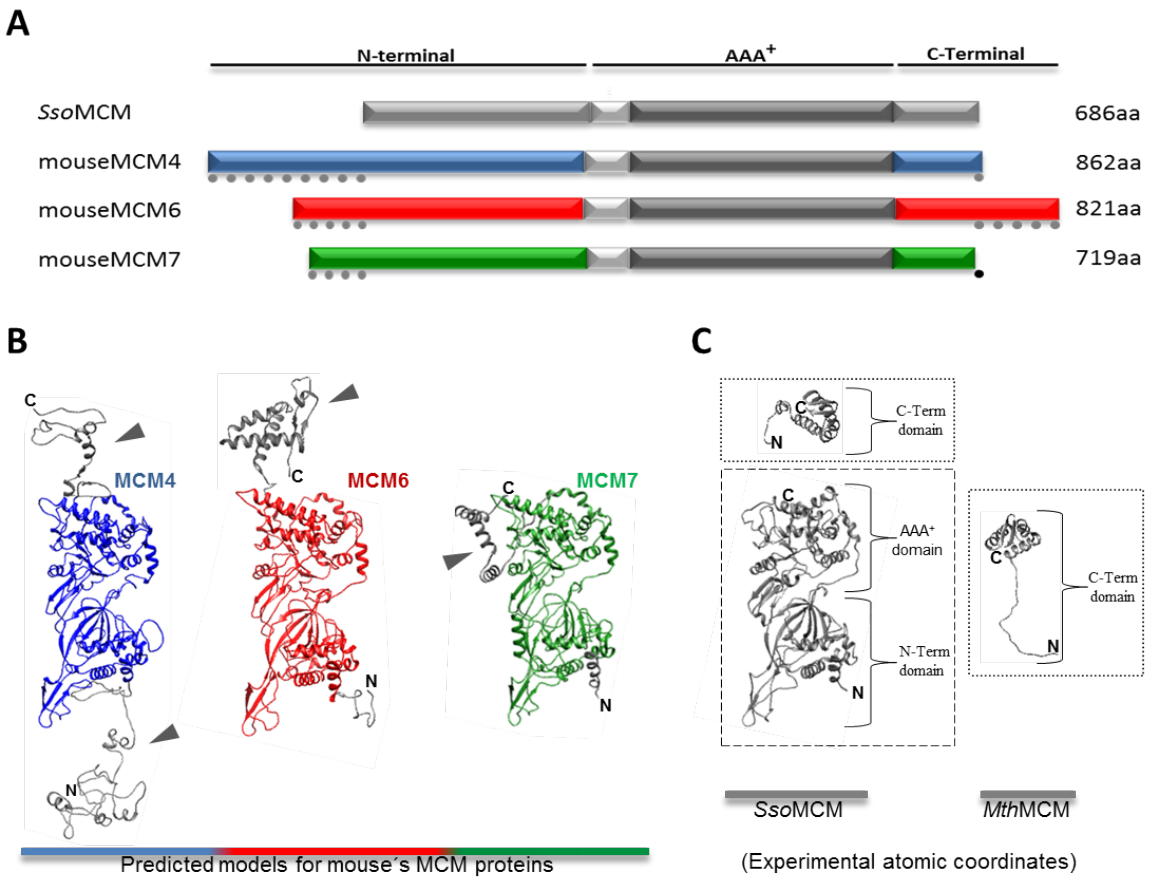


Figure 33. General structure of MCM4,6 and 7. (A) Comparative schematics of *Sso*MCM and mouse MCM4,6 and 7 proteins. Gray dots demark the terminal extensions of the mouse proteins when compared against the archaea MCM, while black dots represent the length difference between the C-terminal region from both MCM7 and *Sso*MCM, being the latter the longest one. **(B)** I-TASSER predicted structures for mouse MCM4 (blue), MCM6 (red) and MCM7 (green) full-length proteins. Arrowheads point to the grey-colored long N-terminal extension of MCM4 and the C-terminal domains of all the three mouse proteins, being the one of MCM6 the crystallographic structure PDB-2KLQ. **(C)** The near full-length structure of *Sso*MCM (PDB-3F9V) is highlighted by a dashed black box, and the C-terminal structures of *Sso*MCM (PDB-2MA3) and *Mth*MCM (PDB-2M45), are highlighted by doted black boxes.

4.2.2. EM structural characterization of the sample

Helicases are macromolecular motors that employ the energy obtained from ATP hydrolysis to carry out its functions. Conformational changes induced by ATP binding and hydrolysis produce a highly dynamic behavior. It has been reported that in the absence of nucleotide, the MCM4/6/7 trimeric state is prevalent. On the other hand, ATP (and a non-hydrolizable ATP analog) favors the dimerization of the trimer to form the MCM4/6/7x2 hexamer, and an intermediate state of trimmers and hexamers is observed with ADP (Ma et al. 2010). By analyzing the MCM4/6/7 sample

in presence of ATP, we were able to capture and study a number of putative ATP binding/hydrolysis-dependent intermediate complexes. Several of the presented architecture features in this work reveal novel structural details never reported before.

4.2.2.1. Description of sample´s features

A -80°C preserved sample purified in the presence of ATP, containing a homogenous complex with a molecular weight agreeing with a hexameric assembly (You & Masai 2008) and formed by stoichiometric quantities of the proteins MCM4, MCM6 and MCM7 (Figure 34 A), was analyzed by negative-stain EM (Figure 34 B). Reference free 2D image classification on a set of 58,010 manually selected particles revealed a highly dynamic sample, mostly populated by flexible hexameric complexes, although lower oligomers were also observed (Figure 34 C). The above mentioned, as well as other more detailed structural features that will be discussed later in this section, becomes more evident after comparing our experimental 2D classes with projections of a canonical MCM hexameric two-tier ring model (Figure 35). The side view of this canonical model presents a two-tire shape, where the N-terminal domains from all the protomers form a first layer relatively flat, which is topped by a thicker layer formed by their AAA^+ domains. The more flexible regions, corresponding to the C-terminus and the N-terminal extensions present in some eukaryotic MCMs, are not visible in the canonical model. The top view has a ring-like shape with six interconnected lobes around a central pore.

It is known that the presence of nucleotide stimulates the formation of the hexameric complex, while lower oligomeric forms of the protein are obtained in the *apo* state (Ma et al. 2010). On the other hand, eukaryote MCM proteins have DNA-dependent ATPase motifs in their AAA^+ central domains (Ishimi 1997b), (Hu et al. 1993), (Koonin 1993), and ssDNA importantly increases the ATPase activity of MCM4/6/7x2 (Kanter et al. 2008). Studies in *Mth*MCM showed that substantial conformational changes occur during ATP binding and hydrolysis (Sakakibara et al. 2009), and in *Ecu*MCM2-7 it has been described a compaction of the ring upon ATPγS binding (Lyubimov et al. 2012). On the basis of the foregoing, we speculate that different states of the helicase (*apo*, ATP-bound and ATP-hydrolysis intermediates), may give rise to the observed conformational changes.

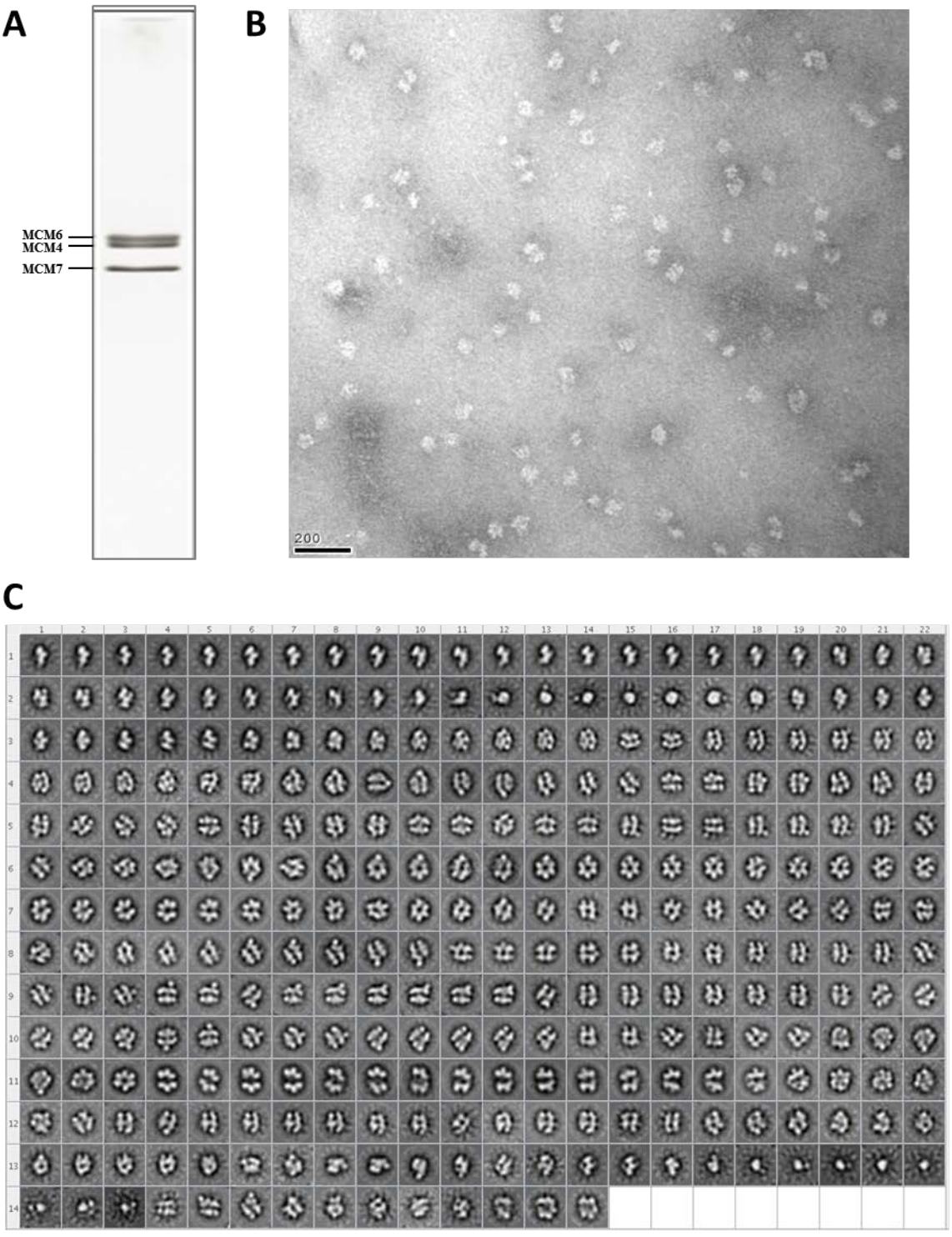


Figure 34. MCM4/6/7 sample and particles. (A) SDS-PAGE gel of a glycerol gradient purified fraction of mouse MCM4/6/7x2 complex. **(B)** Representative negative stain electron micrograph of the sample. **(C)** Reference-free class average classification (CL2D) of the complete set of 58000 particles images of the sample.

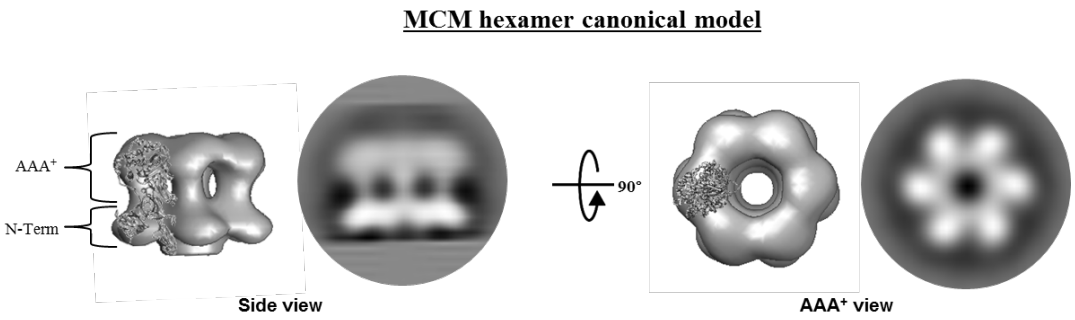


Figure 35. Representative model of the classically described two-tier MCM helicase. Both the AAA$^+$ and side views of the 3D map (*left*) and their corresponding 2D projections (*right*) are show. A monomer of the atomic structure of *Sso*MCM (PDB-3F9V), lacking its C-terminal domain, is fitted into the density model.

2D image analysis showed some novel classes with size and shape compatible with being top/bottom-views of one half of a hexamer only (i.e. a trimer) (Figure 36 A). Other similar classes, but displaying two smaller additional densities located at opposite sides respect to each other, are also observed. Interestingly, a number of images from individual particles revealed the presence of thin, flexible strings emerging from the body of the proposed MCM trimer (Figure 36 B). It is expected that during the averaging of the images, these highly flexible structures get reduced, at best, to the base part, for being the most stable point. Considering the above observations, and the existence of long extensions in some of the MCMs subunits (Figure 33), the additional electron densities in the 2D classes may correspond to the N-terminal and C-terminal sequence extension of MCM4 and MCM6, respectively. Another interesting 2D class average showed what seems to be two separated and confronted trimers. A careful view of the area between the two putative trimers reveals a soft density, and the observation of the single particle images make evident a thin bridge connecting the two trimers (Figure 36 C).

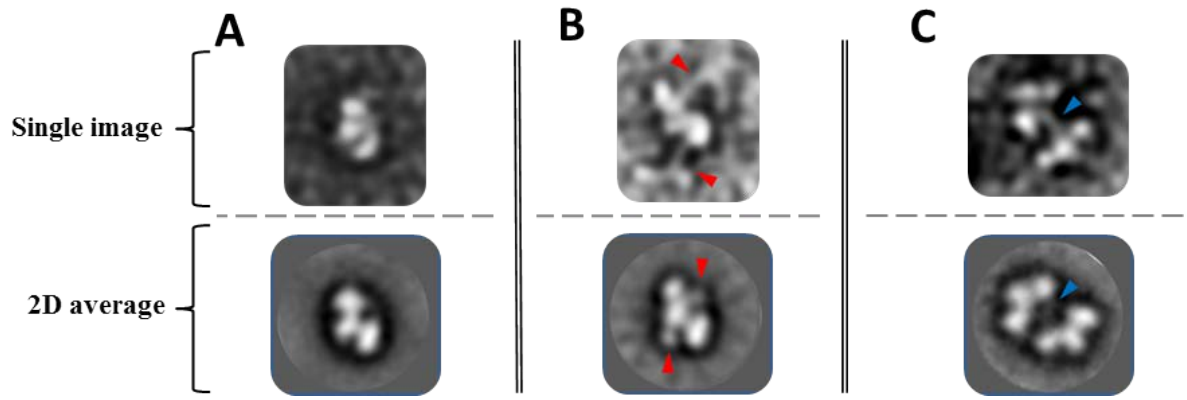


Figure 36. Individual images and reference-free class average of putative top-views from MCM4/6/7. (A) simple trimer, **(B)** trimer with two protruding flexible strings (pointed by red arrowheads) and **(C)** two facing trimers looking as if bound through a thin central bridge (pointed by a blue arrowhead).

In addition to classes similar to the more classically described side and top-view projections of the two-floor ring-like hexameric helicase (Figure 37 A, compared with the projections in Figure 35), and a gapped or even some more widely opened rings (Figure 37 B), we also captured top- and side-views of hexamer conformations showing, quite sharply, the presence of one strong extra density, which, by comparison with the canonical side view of an MCM helicase (Figure 35), seems to be located at the C-terminal region of the ring (Figure 37. C). A different novel conformation is a side view presenting two similar additional electron densities; once again, by comparison with the canonical side view of an MCM helicase (Figure 35), both densities could correspond to the C-terminal domain of some of the protomers (Figure 37 D). These two densities are less strong than the single one observed in the aforementioned 2D classes, reflecting their smaller size. KenDerSOM (Smoothly Distributed Kernel Probability Density Estimator Self Organizing Map) analysis was performed in order to confirm the presence of the two different densities, by discarding the possibility that average of particles with only one extra mass, but in different positions, were aligned, creating an artifactual class that appeared to have two extra densities (Figure 38). Two copies of MCM6 form part of the hexameric MCM4/6/7x2, and this protein has a substantial C-terminal extension when compared with both MCM4 and MCM7. Then, following this reasoning, it seems very plausible that each of the two extra masses protruding from the AAA^+ region of the helicase may correspond to the HTH domain of both MCM6 protomers. In the same line, and considering that in the classes showing only one extra mass emerging from the AAA^+ this is much more electron dense than the ones in class with two masses, one might think that the

strong electron density could be formed by the partial or complete interaction of the HTH motifs of both MCM6s, which in turn would support the proposed model where the two MCM6s are contiguous (Figure 10 B). Finally, we observed a minority group of small particles that could, tentatively, correspond with views of free monomers (Figure 37 E), which are around 80-95 kDa.

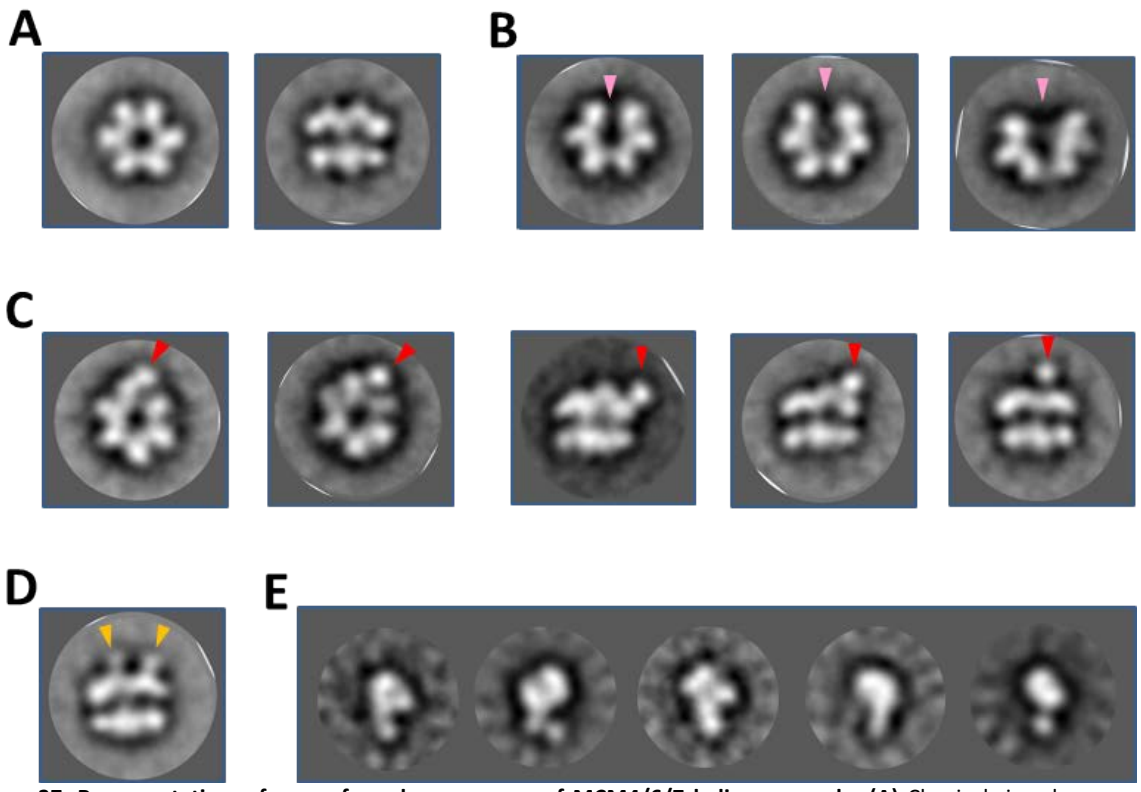


Figure 37. Representative reference-free class averages of MCM4/6/7 helicase sample. (A) Classical ring-shape and side views of and MCM helicase; **(B)** open rings with different level of aperture (pink arrowhead points to the gap); **(C)** ring-shaped (first two *left* panels) and side views (last three *right* panels) of the helicase, displaying a sharp extra electron density (pointed by a red arrowhead); **(D)** a side view with two light extra electron densities (pointed by yellow arrowheads) and **(E)** putative side views of monomers.

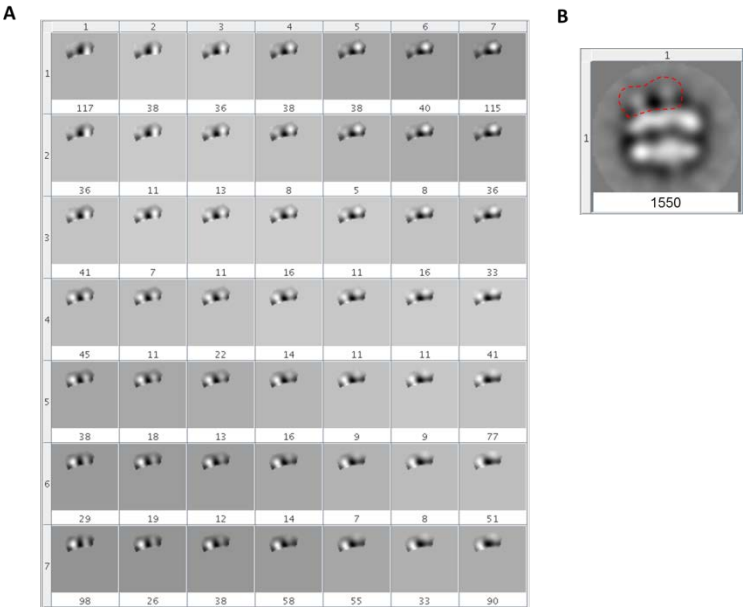


Figure 38. KenDerSOM analysis. (A) Analysis of a zone showing two extra densities (demarcated by a red dashed contour) in **(B)** a reference-free class average. Numbers refer to the particles on each class.

Interestingly, some helicase pairs can be found bound to each other through a thin cord (Figure 39 A), suggesting that protruding extensions are involved in the interaction between helicase particles. Additionally, some string-like structures formed by a variable number of helicase particles were also present (Figure 39 B).

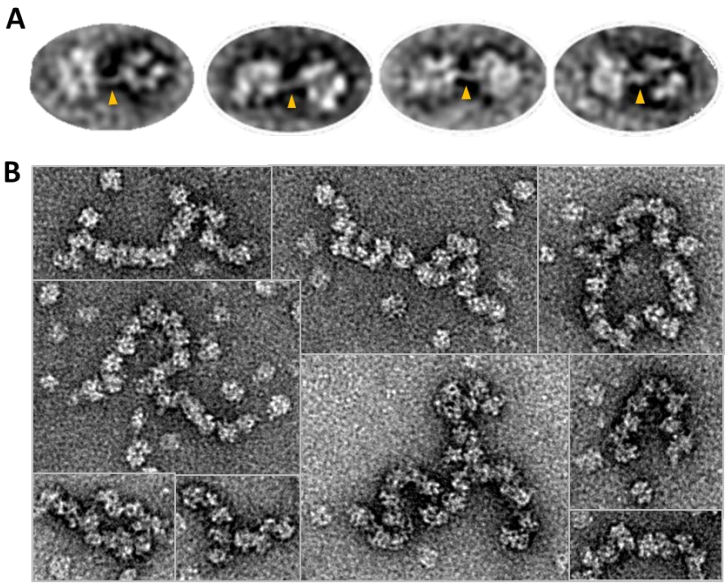


Figure 39. Closely interacting helicase particles. Representative images of **(A)** helicase pairs interconnected by a fine thread in between (pointed by orange arrowheads), and **(B)** strings formed by helicase particles.

4.2.2.2. 3D reconstructions and fitting of atomic structures

To construct active ATPase sites within the helicase, a proper alignment between subunits must occur; in this way, the correct *cis* and *trans* interactions of the catalytic motifs at the interfaces of the proteins are formed (Figure 9) (Brewster et al. 2008), (Miller et al. 2014). The intersubunit nature of the active sites, in conjunction with the proposed protomer arrangement of MCM4/6/7x2 (Figure 10 B), imply that, despite having the same protomer composition, the two trimers do not have symmetric intersubunit interactions. That is, in the hexamer there would be both a MCM7→4→6 and a MCM6'→4'→7' trimer (hereinafter an apostrophe sign will be used to differentiate the two subunits of the same type of protein for each of the two trimers), where the subunits to the left of the arrow provide the motifs in *trans*, while the subunits to the right contribute with the motifs in *cis*. The organization of the MCM7→4→6 trimer has normally been considered to be the same than the one that is contained in the helicase MCM2-7(Figure 10). However, because the two trimers are not equally organized, their structures do not necessarily be exactly the same, neither their functions within the complex. Indeed, all MCM4/6/7x2 structures presented in this work show a clear symmetry mismatch. An important additional consideration is the fact that ATP binding promotes the stabilization of the hexamer, while ATP hydrolysis has a destabilization effect that shift the equilibrium toward the trimer (Ma et al. 2010).

Using negative-stain EM, single-particle classification and reconstruction methods, we obtained multiple 3D maps corresponding to different oligomeric and conformational states of mouse MCM4/6/7. In order to facilitate the reader to recognize the side and front views within our 3D maps, as well as the three principal structural regions (N-terminus, AAA^+ and C-terminus), we have adopted a similar lay out as the one in Figure 7 C. Importantly, it must be noted that fitting of atomic structures within the presented low resolution density maps is only a rough guide of their likely relative positions, and does not intend to represent the absolute orientation in any of the cases.

Identification of isolated MCM trimer configurations

To date, a number of biochemical works have reported the formation of a MCM4/6/7 heterotrimer (Musahl et al. 1995), (Sherman et al. 1998), (Ma et al. 2010), (Ichinose 1996), (Ishimi 1997a), but there is not further information at the structural level. Along the process of classification and refinement of the different 3D maps reconstructed from our analyzed EM sample data, some of the obtained structures presented a principal body with shape and size compatible with the mass of a MCM trimer. Some structural variants of this putative trimer were observed; an apparently more compact state as well as conformations showing two extended and flexible protrusions emerging from the principal body. Hence, we present a representative structure of each of these putative states. A first 28 Å estimated resolution 3D map shows three column-like structures bound together by their top and bottom parts, creating a complex that resembles the side-view of the hexameric helicase (Figure 40). Three copies of the crystallographic structure of the near full-length *Sso*MCM (PDB-3F9V) fits well on this volume, which shows, unequivocally, that this is a MCM trimer. An additional feature is an extra density located at the interface between two of the subunits, protruding above what could correspond to their AAA^+ region. Indeed, it is known that MCM6 contains a C-terminal domain much bigger than the ones of MCM4 and MCM7, and we noted that the atomic coordinates of the HTH C-terminal structure of *Hs*MCM6 (PDB-2KLQ) can be nicely accommodated in this extra space.

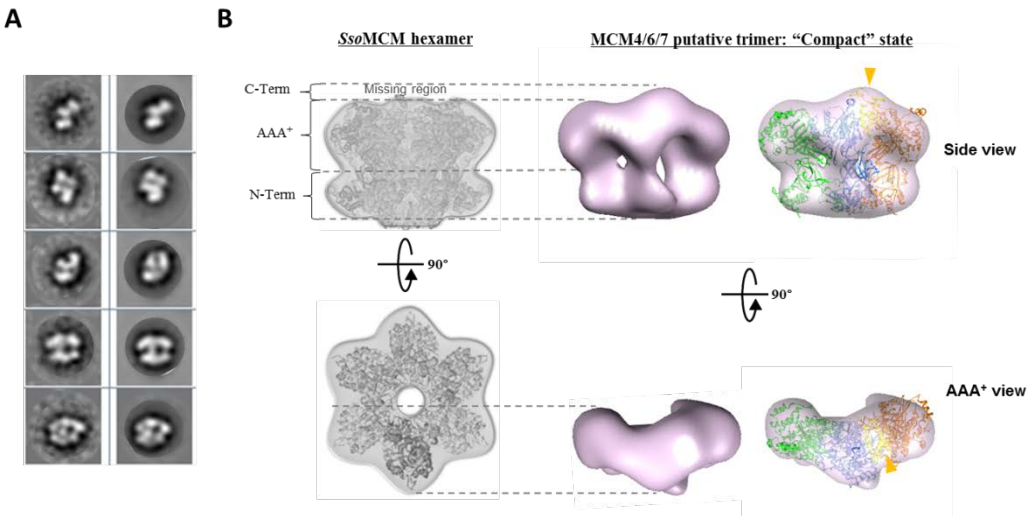


Figure 40. Putative MCM4/6/7 trimer reconstruction of a likely compact conformation. Designation of the C-terminal AAA$^+$ and N-terminal domain regions has been done by comparison with a *Sso*MCM hexamer model. **(A)** Reference-free class averages (*left* column) and the corresponding forward projections (*right* column) of the 3D model. **(B)** From left to right there are shown the side views (*top* panels) and AAA$^+$ views (*bottom* panels) of: an *in silico* assembled hexamer model of *Sso*MCM lacking the C-terminal region (modified from (Brewster et al. 2008)); a solid colored 3D map of a putative MCM4/6/7 trimer and the fitting within the same map of three copies of the *Sso*MCM monomer (PDB-3F9V, in light-green, light-blue and orange) and the C-terminal domain of *hs*MCM6 (PDK-2KLQ, in yellow). Yellow arrowhead points to the extra density where the C-terminal structure has been fitted.

The second map, resolved at an estimated 26 Å resolution, also shows dimensions that are congruent with a possible MCM trimer, although at lower threshold levels features two additional, extended electron densities (Figure 41). One of these densities extrudes from the putative AAA$^+$ region of one of the side protomers and projects to the inner part of the trimer. A second extra density is projected to the external side of the trimer and departs from the putative N-terminal region, at the interface between the subunit with the putative AAA$^+$ extension and the protomer at the middle of the trimer. By their position, these protrusions could correspond to extended versions of the C-terminal of MCM6 and the N-terminal extension of MCM4. 2D averages and single particle images displaying small and flexible extra densities, respectively, suggest that these projections are not artifacts of the 3D reconstruction. Finally, there is a small extra mass protruding from the interface between the central protomer and the one opposite to the firstly described two extended densities; in this space the C-terminal structure of *Mth*MCM (PDB-2MA3) can be fitted. Due to the expected flexibility of the projections and the limited resolution of the maps, it is not possible to ensure if the differences between the maps of the two trimers are just variations of the same complex, or if they really point to structural characteristics inherent to each of the two putative trimers, MCM7→4→6 and MCM6'→4'→7'.

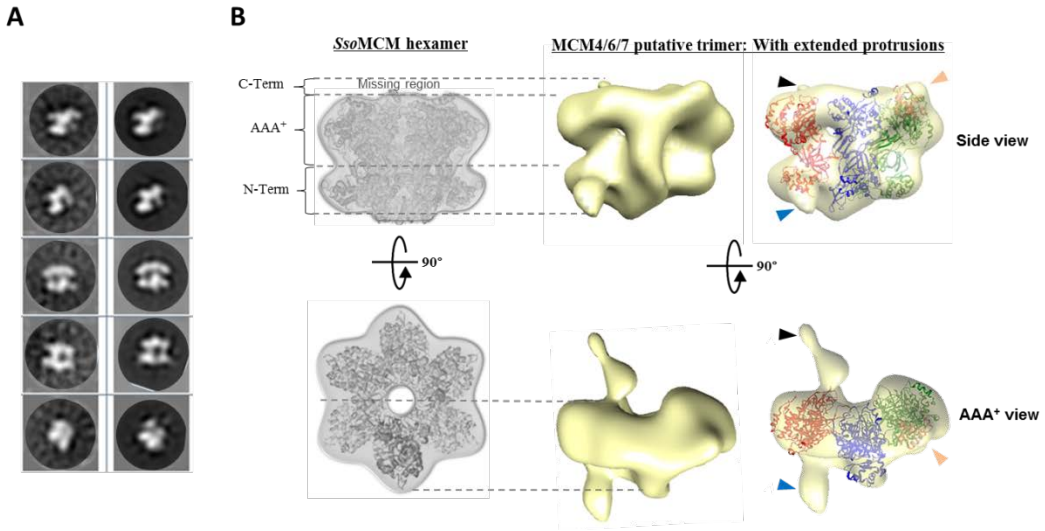


Figure 41. Putative MCM4/6/7 trimer reconstruction showing extended densities. Designation of the C-terminal AAA$^+$ and N-terminal domain regions has been done by comparison with a *Sso*MCM hexamer model. **(A)** Reference-free class averages (*left* column) and the corresponding forward projections (*right* column) of the 3D model. **(B)** From left to right there are shown the side views (*top* panels) and AAA$^+$ views (*bottom* panels) of: an *in silico* assembled hexamer model of *Sso*MCM lacking the C-terminal region (modified from (Brewster et al. 2008)); a solid colored 3D map of a putative MCM4/6/7 trimer and the fitting within the same map of three copies of the *Sso*MCM monomer (PDB-3F9V, in red, blue and green) and the C-terminal of *Sso*MCM (PDB-2M45, in pale-pink). Pale-pink arrowhead points to the area where the C-terminal atomic coordinates have been fitted .Extra densities that would putatively corresponded to HTH motif of MCM6 and N-terminal extension of MCM4, are pointed with black and blue arrowheads, respectively.

Incipient dimerization of MCM trimers

Mutational and biochemical studies on human MCM4 have implicated residues at its N-terminal domain as essential for dimerization of MCM4/6/7 trimers (You et al. 2002). Along this line, one of our 3D maps, resolved at an estimated resolution of 30 Å, presents two principal bodies, each one likely to be a MCM trimer, that are connected through a bridge emerging at one side of the two putative trimers, near the protomer of the middle (Figure 42), which could correspond to MCM4 (Figure 10). After fitting six copies of *Sso*MCM monomer (PDB-3F9V), it remains empty space within the bridge, which would be putatively filled by the interacting N-terminal extensions of the MCM4 subunits. Additionally, both trimers had an extra density extruding at the opposite site of the bridge, near to the subunit of the middle. Each of these densities has enough room to accommodate the atomic coordinates of the C-terminal structure of *hs*MCM6 (PDB-2KLQ).

This 3D reconstruction would provide direct visual evidence in support of the oligomerization role of the N-terminal of MCM4, which is involved in the assembly of two MCM trimers into the

heterohexameric complex MCM4/6/7x2. However, further labeling studies for mapping the N-terminus of MCM4 are be required to confirm the identity of the observed bridge.

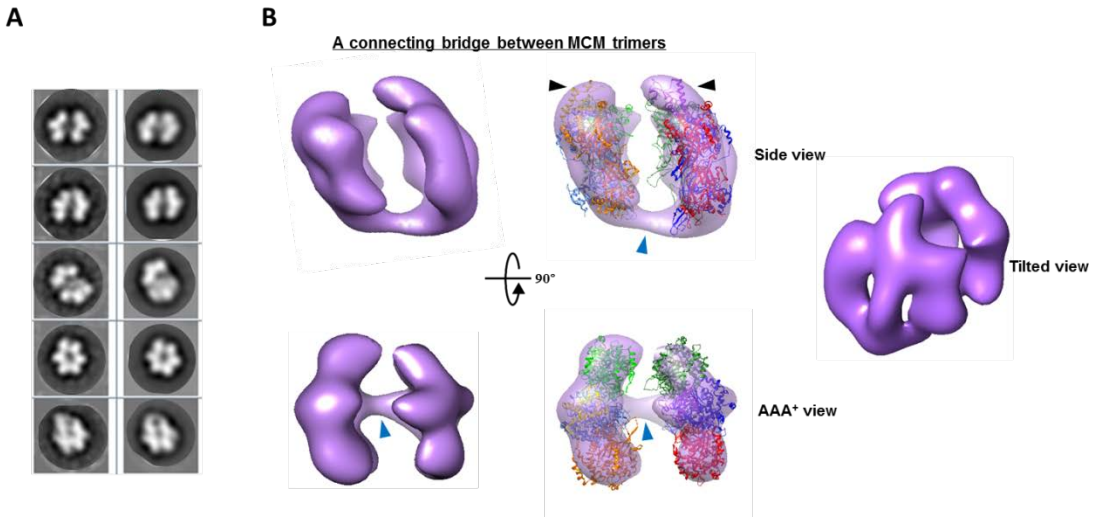


Figure 42. Early stage on dimerization of MCM trimers. (A) Reference-free class averages (*left* column) and matching forward projections (*right* column) of the density map. **(B)** Putative side (*top* panels) and AAA$^+$ (bottom panels) views of the 3D map and the respective fitting of six copies of the *Sso*MCM monomer (PDB-3F9V, in orange, light-blue, light-green, red, blue and green). Blue arrowheads point to an extra density that connects the two trimers. Black arrowheads point to two extra densities at the putative C-terminal region, each with a copy of the HTH motif of *Hs*MCM6 (PDB-2KLQ, in light-orange and purple) fitted inside. An intermediate tilted view of the map is shown to get a better idea of its overall shape.

Architecture of the heterohexamer MCM4/6/7x2: Structures of multiple conformational classes exemplifying the flexibility of the complex

Hexameric MCM helicases have been shown to be highly dynamic molecular machines, able to make conformational changes implying different movements of their N-terminal, AAA$^+$ and C-terminal regions, as well as partial or complete disruption of the binding interface between some of the subunits.

A number of 3D maps, with estimated resolutions between 28-33 Å (Table 2), corresponding to different structural classes adopted by the MCM4/6/7x2 heterohexamer, were classified and refined as described in Materials and Methods (Figure 16). Common to all classes is an almost

planar configuration that can be divided into three regions, one on top of the other, containing the N-terminal (at the bottom), the AAA^+ (at the middle), and the protruding C-terminal (at the top) domains of some of the protomers. Six copies of the near full-length crystal structure of *Sso*MCM (PDB-3F9V) can be fitted as rigid bodies into each of the 3D EM maps; some densities left unoccupied, may correspond to N-terminal or C-terminal extensions present in some eukaryotic MCM´s. Consistently with the dimensions showed by an archaea MCM hexamer model lacking its C-terminal region (side view in Figure 7 C and Figure 35), the N-terminal region of our maps is less voluminous than the AAA^+ region, however, additional masses at the putative C-terminal region of some protomers, are systematically observed. In all cases, an asymmetrical top and bottom opening of the central helicase channel is observed, being the one at the AAA^+ region the wider. Up to now, works on MCM2-7, the proposed replicative helicase, suggest that it depends on additional proteins or special buffer conditions for both helping to seal its MCM2/5 gate and acquire a planar configuration which, respectively, seem to be possible mechanisms for avoiding DNA to slide out through the ring gap and to get a correct alignment of the intersubunit active sites in order to activate the helicase activity of the complex.

To gain further insights into MCM4/6/7x2 organization, we made a comparative analysis of the structural features between our 3D reconstructions and a number of reported MCM2-7 structures. As a group, our set of EM maps is a revealing presentation of mostly planar conformations of closed, open or nicked structures that, analyzed together, suggest a number of changes involving the movement and interaction of the putative C-terminal domains of two consecutive protomers, as well as the disruption and creation of new contacts between different subunits. In the following we will analyze each of these classes.

Map *Class*	Estimated resolution (FSC 0.5 cutoff)
1	27 Å
2	33 Å
3	32 Å
4	31 Å
5	30 Å
6	28 Å

Table 2. Estimated resolution of the 3D maps of MCM4/6/7x2 hexamers Class 1 to 6, by the Fourier Shell Correlation (FSC) criteria at a value of 0.5.

Class 1: A closed ring structure on which the proposed architecture of the hexamer is presented

A first MCM4/6/7x2 map, corresponding to *Class1*, presents the general shape of a two-tiered ring (Figure 43). The N-terminal opening of the central channel forms a pore with a diameter of around 25Å, and the aperture toward the AAA^+ region is around twice in size (diameter of around 60 Å x 45 Å). Two extra masses protruding at different points from the AAA^+ region of the map stand out; a small wedge and a large lump (densities displayed in dark blue and cyan, respectively, in Figure 43 B).

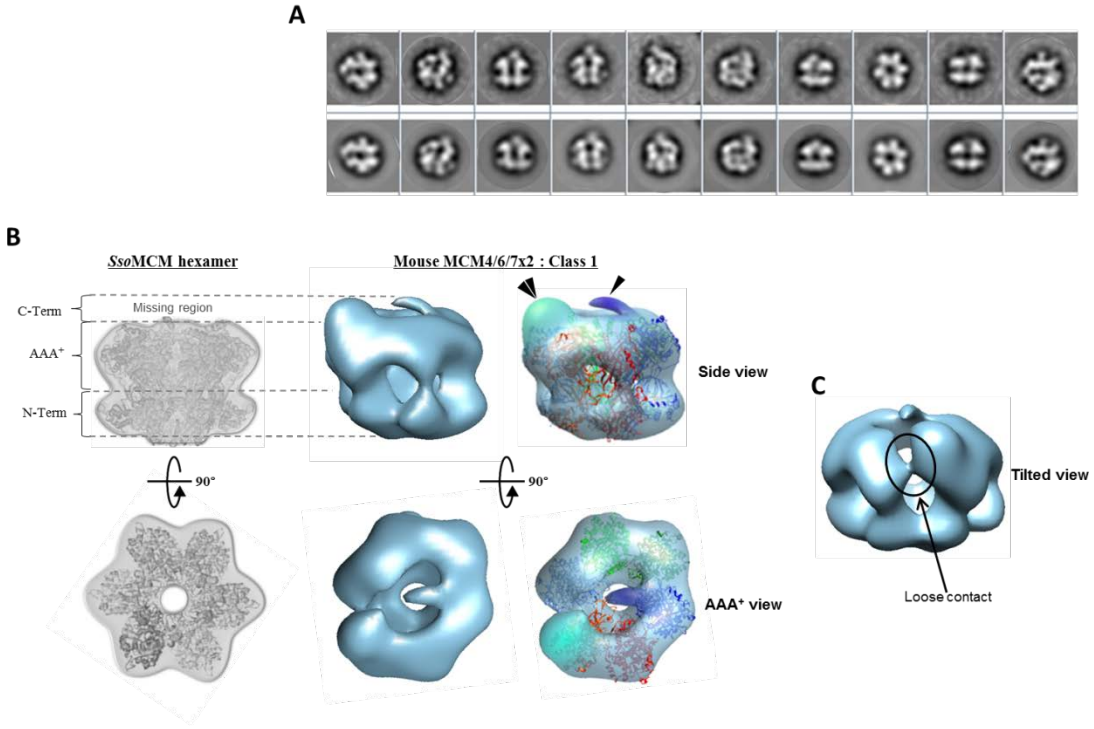


Figure 43. *Class1*: A MCM4/6/7x2 closed ring displaying two different extra masses, a big and a small one. (A) Reference-free class averages (*first* row) and the corresponding forward projections (*second* row) of the 3D map. **(B)** From left to right there are shown the side views (*top* panels) and AAA^+ views (*bottom* panels) of: an *in silico* assembled hexamer model of *Sso*MCM lacking the C-terminal region (modified from (Brewster et al. 2008)); a solid colored 3D map of MCM4/6/7x2 model *Class 1*, and the fitting within the same map of six copies of *Sso*MCM monomer (PDB-3F9V), where two additional empty densities at the C-terminal region, are highlighted in cyan and dark blue, and pointed by a double or single arrowhead, respectively. **(C)** Tilted view of the map, showing a point of weak contact observed between two of the protomers.

The first mass we want to draw the attention to is the smaller density, which is clearly visible both in front and side views (density displayed in dark blue in Figure 44 A). This arched finger-like structure, which is subtly turned against the bigger extra mass, extends toward the opening of the

central channel at the AAA$^+$ region, resulting in a spatial disruption that creates two slightly asymmetric invaginations of around 20 Å each, which is still big enough to accommodate ssDNA.

The length of the amino acid sequence from the HTH motif of both *Mth*M*CM* (PDB-2MA3) and *Sso*M*CM* (PDB-2M45), is similar to the one of the C-terminus of mouse MCM4 (Figure 32 and Figure 33), and we observed that the atomic coordinates from PDB-2MA3 fitted reasonably well into the small extra density of our map. In addition to the small extra density (a putative HTH motif) described above, there is also a big protrusion emerging at the C-terminal region (Figure 44 B), which makes a loose contact with one of its neighbouring subunits (Figure 43 C). We observe that this mass is large enough to contain two copies of the C-terminal structure of *Hs*MC*M6* (PBD-2KLQ).

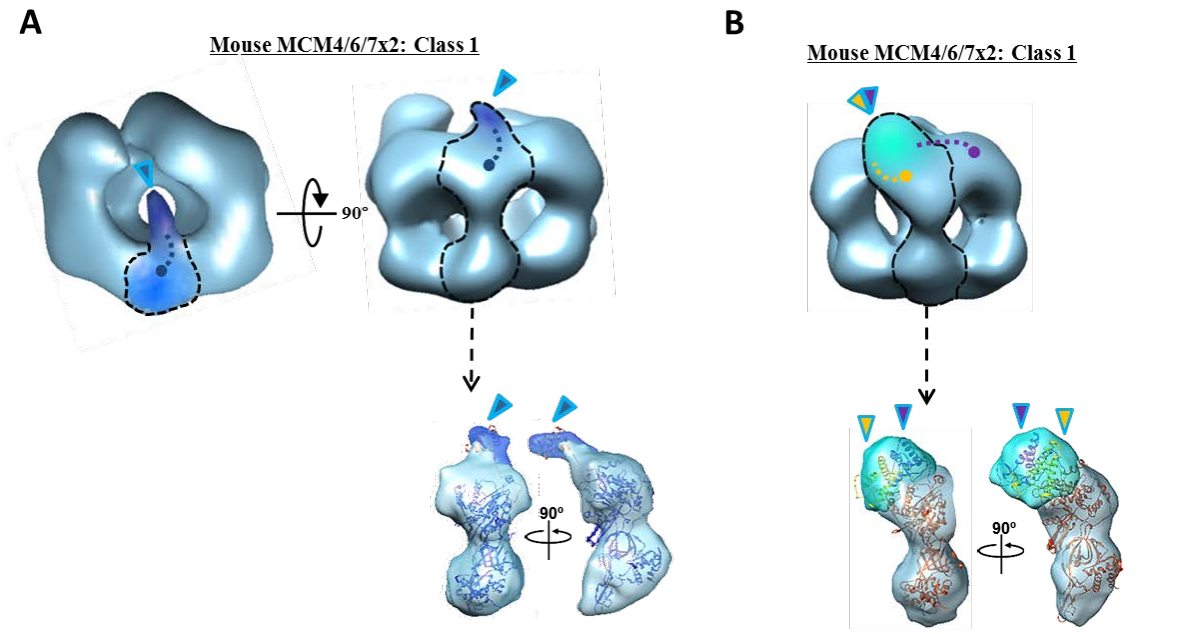


Figure 44. Fitting of atomic structures within two extra masses protruding from MCM4/6/7x2 map *Class 1*. (A) AAA$^+$ and side views of the map, where a black dashed contour highlights a protomer with a small extra density (dark blue-blushed area pointed by a cyan-contour blue arrowhead). Blue doted-line represents a symbolic linkage between the AAA$^+$ region of a subunit and its putative HTH domain. Two different side views of the segmented subunit are shown, with the structure of *Sso*MCM (PDB-3F9V) and C-terminus of *Mth*MCM (PDB-2MA3) fitted inside. **(B)** Side view of the map, where a black dashed contour highlights a protomer with a big extra density (cyan-blushed area pointed by a cyan-contour double-colored yellow/purple arrowhead). Yellow and purple doted-lines represent a symbolic linkage between the AAA$^+$ region and the respective putative HTH domains of two consecutive subunitssubunits. Two different side views of the segmented subunit are shown, with the structure of *Sso*MCM (PDB-3F9V) and two copies of HTH *Hs*MCM motif (PDB-2KLQ) fitted inside.

Interestingly, the features above-mentioned have, in some way, been partially described in previous EM maps of different MCM2-7 helicase complexes (Figure 45), a fact that has provided us with grounds to propose a tentative structural comparison between our current *Class 1* map and those other previously reported ones.

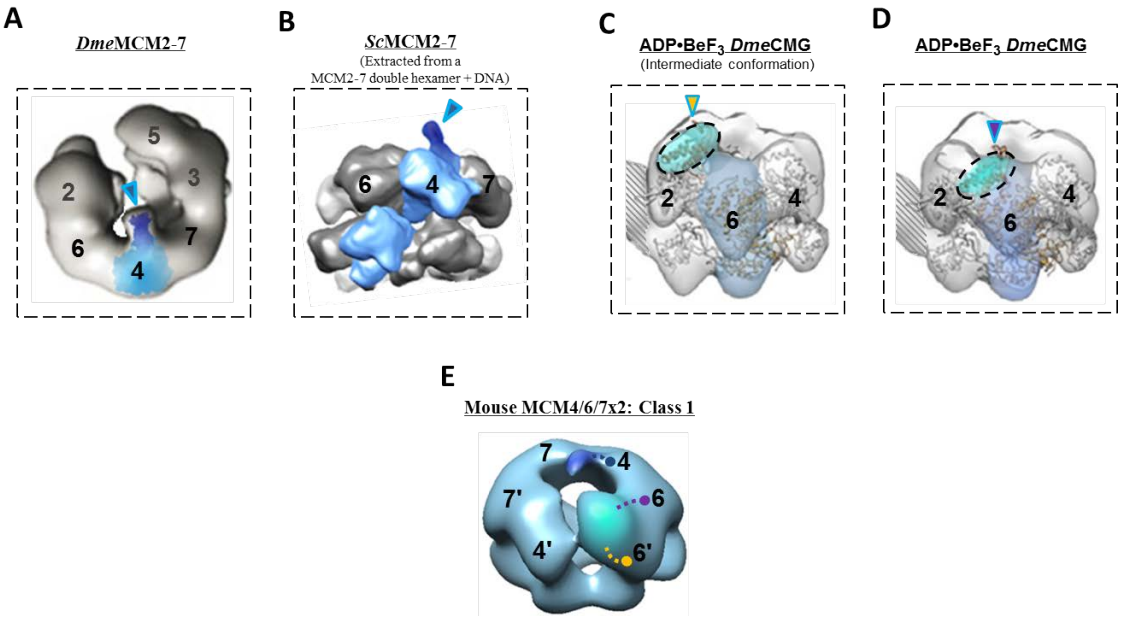


Figure 45. Proposed architecture of MCM4/6/7x2 map *Class1*, based on structural similarities with a number of MCM2-7 EM structures featuring some comparable extra masses. Structure of **(A)** *Dme*MCM2-7 open ring (figure adapted from (Lyubimov et al. 2012)) and **(B)** *Sc*MCM2-7 hexamer (extracted from a MCM2-7 double hexamer + DNA complex, EMD-5857), where cyan-contour blue arrowheads point to and small density (dark blue-blushed area) protruding from MCM4. **(C)** Structure of an alternate conformation of ADP•Be_3 bound *Dme*CMG (Figure adapted from (Costa et al. 2011)), where cyan-contour yellow arrowhead points to the cyan-blushed area that is partially over MCM2, which have the C-terminal structure of *Hs*MCM6 (PDB-2KLQ) fitted. **(D)** Structure of ADP•Be_3 bound *Dme*CMG, where cyan-contour purple arrowhead points to the cyan-blushed area at the MCM2/6 interface with the C-terminal structure of *Hs*MCM6 (PDB-2KLQ) fitted (Figure adapted from (Costa et al. 2011)). **(E)** Designation of all the subunits within mouse MCM4/6/7x2 map *Class1*, where each, blue, purple and yellow doted-lines represent a symbolic linkage between the AAA^+ region of a subunit and its putative HTH domain. For all cases, numbers represent the different subunits. Black-dashed boxes enclose structures resolved in studies other than the present one.

Indeed, similar to our observed small extra mass, in a report showing a EM map of DmeMCM2-7 (structure not deposited in the Electron Microscopy Data Bank, EMDB), a small wedge-like structure is observed arising from the AAA+ domain of MCM4 (Figure 45 A) (Lyubimov et al. 2012). Furthermore, in a ssDNA loaded ScMCM2-7 double-hexamer complex (EMD-5857), a small and elongated extra density extruding from MCM4 of both rings is observed, although in this case it extends upwards (Figure 45 B) (Sun et al. 2014). Perhaps this change of orientation is due to a stearic exclusion produced by the presence of the ssDNA, or as a result of the general conformation acquired by the two interlocked and misaligned hexamers. It must be underlined

that in the above two mentioned MCM2-7 structures, this small extra mass is slightly bent in direction to MCM6. All the above observed similarities allowed us to propose that the small protrusion may correspond to the HTH domain of MCM4, so that it can be used to localize the position of this subunit in our map and, at the same time, suggest the position of its neighbouring subunits, MCM6 and MCM7 (Figure 45 E).

On the other hand, following the line of the proposed fitting of two copies of the HTH MCM6 motif inside the big extra mass in our map Class1, there are some works that give support and insights about this idea. Given the putative position of MCM4 (identified by the small extra mass), the location of the lump and the model of the subunit distribution of MCM4/6/7x2 (Figure 10B), all these observations could imply that the subunit MCM6 besides the MCM4 with the finger-like protrusion, extends its C-terminus all over MCM6' so the HTH parts of both, MCM6and MCM6', can interact. A structural study on the Drosophila CMG complex would support similar positional shifts of the winged helix domain of MCM6 (Costa et al. 2011). Upon binding of ADP•BeF3 by DmeCMG complex (structure not deposited in EMDB), an intermediate configuration of the DmeMCM2-7 ring has a MCM6 subunit with a putative extra mass extruding from its AAA+ domain that is partially over MCM2 (Figure 45 C). In an alternative conformation of Drosophila ADP•BeF3 CMG complex (EMD-1832), the putative HTH domain of MCM6 seems to be displaced over the MCM2/6 interface, apparently completely recessed against the AAA+ region (Figure 45 D). Making use of all these observations we propose that in our presented volume of the closed ring, the interacting HTH motifs of the MCM6-6' protomers would form the big extra electron density. In view of all the foregoing, having putatively determined the position of MCM6, MCM6' and MCM4, and following the model proposed by previous biochemical studies (Yabuta et al. 2003), (Yu et al. 2004), (Xu et al. 2013) (Figure 10B), the remaining subunits can also be putatively positioned (Figure 45 E).

Class 2: A notched ring showing two extensions extruding from the AAA$^+$ region

3D map *Class 2* (Figure 46) reveals the conformation of a ring with a discontinuity at the AAA$^+$ region, which is flanked by two protomers displaying a prominent mass at their respective C-terminus. Indeed, in spite of their different shapes, one molecule of the HTH motif of *Hs*MCM6 (PDB-3F9V) can be fitted inside each of these extensions; hence these protuberances are assigned

to the HTH motif of each of the two MCM6 protomers in the complex. Considering that these protuberances are flanking the notch, we propose that the two protomers at the nick interface of this map may be the copies of the protein MCM6. The protomer presenting the most extended C-terminus looks very loosely bound by its N-terminal part to the neighbouring subunit at the contrary side of the nick (Figure 46 C). This somehow apparently unstable architecture could perhaps be interpreted as an indicative of being a transitional state that moves toward a more stable conformation. Having putatively assigned the positions of the two MCM6s, and following the proposed model organization of MCM4/6/7x2 (Figure 10 B and Figure 45 E), we can complete the assignation of the remaining subunits.

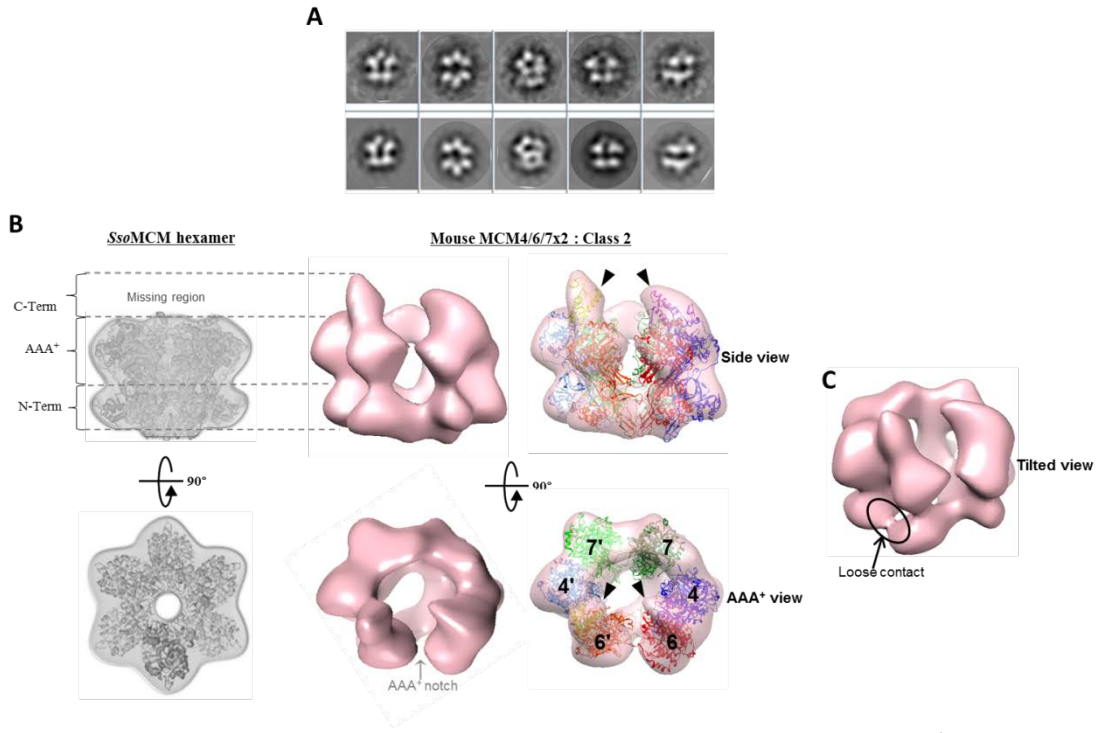


Figure 46. *Class 2*: A MCM4/6/7x2 complex with two densities protruding from a notched AAA$^+$ region. (A) Reference-free class averages (first row) and the corresponding forward projections (second row) of the 3D map. **(B)** From left to right there are shown the side views (*top* panels) and AAA$^+$ views (*bottom* panels) of: an *in silico* assembled hexamer model of *Sso*MCM lacking the C-terminal region (modified from (Brewster et al. 2008)); a solid colored 3D map of MCM4/6/7x2 model *Class 2*, and the fitting within the same map of six copies of *Sso*MCM monomer (PDB-3F9V, in orange, light-blue, light-green, red, blue and green). Black arrowheads point to two extra densities at the C-terminal region, each with a copy of the HTH motif of *Hs*MCM6 (PDB-2KLQ, in light-orange and purple) fitted inside. **(C)** Tilted view of the map showing a point of weak contact between the N-terminal parts of two protomers. Numbers represent the putative organization of the different subunits.

Class 3: An open ring with a C-terminal protrusion that partially occludes the central pore

The structure of an almost planar, open ring, is observed in *Class 3* map (Figure 47 A and B). We draw the attention to both protomers at the gap, which are notably longer than the rest at the C-terminal region, letting enough space for the HTH structure of *Hs*MCM6 (PDB-2KLQ) to be fitted in addition to the respective copy of the almost full-length *Sso*MCM monomer (PDB-3F9V). Following a similar reasoning used in the analysis of map *Class 2*, we propose that the opening of this volume occurs at the MCM6'-6 interface, allowing us to design the position of the remaining subunits. The putative C-terminus of MCM6' looks almost fused with the AAA^{+} domain, suggesting that the HTH motif could be in a recessed position. On the other hand, the putative MCM6 monomer is a little displaced upward in relation to the horizontal axis of the bottom wheel, and its C-terminal part looks like a lobe projected toward the central channel of the ring, so that it is partially obstructing this pore. Similar C-terminal structures can be recognized in some MCM helicase density maps from other organism. For example, an extended density at the C-terminus of MCM6 is observed, at low contour levels, in the structure of an apo *Dme*CMG complex (Figure 47 C) (Costa et al. 2011). MCM´s HTH domains positioned against the central channel had been also visualized in one of the rings of the *Mth*MCM double hexamer, in presence of dsDNA (Costa et al. 2006a). Furthermore, the recently solved structure of a *Hs*MCM2-7 + DNA complex shows one prominent C-terminal extension (Hesketh et al. 2015).

In a study of the C-terminal winged helix fold domain of *hs*MCM6, this region, by itself, failed to bind DNA (Wei et al. 2010). An enhanced affinity for ssDNA and dsDNA was observed in a *Mth*MCM C-terminal deletion mutant, suggesting that the HTH motif might be implicated in controling DNA accessibility to the helicase pore (Jenkinson & Chong 2006). The aforementioned observations let us think that perhaps, the putative HTH MCM6 that extend toward the central channel, in the *Class 3* map, could be a flexible structure acting as a transient stearic impediment for DNA to go into the central channel of the helicase under certain conformation of the hexamer.

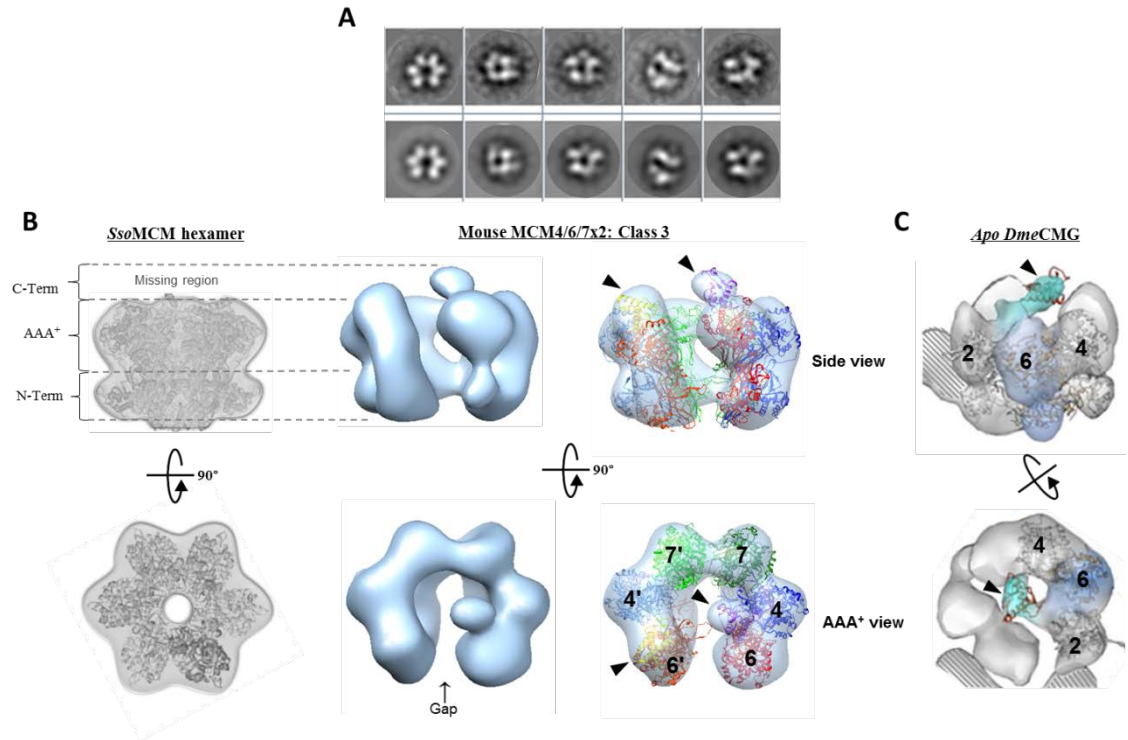


Figure 47. *Class 3*: A MCM4/6/7x2 open ring with two longer subunits at the gap interface. (A) Reference-free class averages (first row) and the corresponding forward projections (second row) of the 3D map. **(B)** From left to right there are shown the side views (*top* panels) and AAA$^+$ views (*bottom* panels) of: an *in silico* assembled hexamer model of *Sso*MCM lacking the C-terminal region (modified from (Brewster et al. 2008)); a solid colored 3D map of MCM4/6/7x2 model *Class 3*, and the fitting within the same map of six copies of *Sso*MCM monomer (PDB-3F9V, in orange, light-blue, light-green, red, blue and green). Black arrowhead points to an extra density at the putative C-terminal region of the complex, where the HTH motif from *Hs*MCM6 (PDB-2KLQ, in purple) is fitted. **(C)** Tilted and AAA$^+$ views of *apo Dme*CMG (EMD-1833), with models of MCM subunits 2. 4 and 6, and the *Hs*MCM6 HTH structure (PDB-2KLQ) fitted inside. The density corresponding to MCM6 is highlighted with blue blush and the one containing the atomic coordinates of human HTH MCM6 motif (PDB-2KLQ), pointed by a black arrowhead, is highlighted with cyan blush. (Figure adapted from (Costa et al. 2011)). For all cases, numbers represent the different subunits.

Class 4: A loosely closed ring with two interacting C-terminal protrusions

The toroidal structure of map *Class 4* (Figure 48) presents a highlighting feature at the connecting interface between two of the subunits, which is looser than in the rest, both at the bottom and at the top. At this interface, the link of the upper part seems to be formed between the C-terminus of the two flanking protomers, contrasting with other pair interaction between proteins, which make contact at the level of their AAA$^+$ domains. The fact that two copies of the HTH motif of *Hs*MCM6 (PDB-3F9V) can be fitted along this C-terminal connection, suggest that the loose interface is formed by the putative MCM6'-6 proteins. This putative HTH-HTH connection goes in the same line of our idea that the big lump presents in *Class 1* map could be formed by the interacting HTH domains of both MCM6. After defining the putative position of the MCM6s

protomers, the rest of the subunits were organized according with the model presented in Figure 10 B.

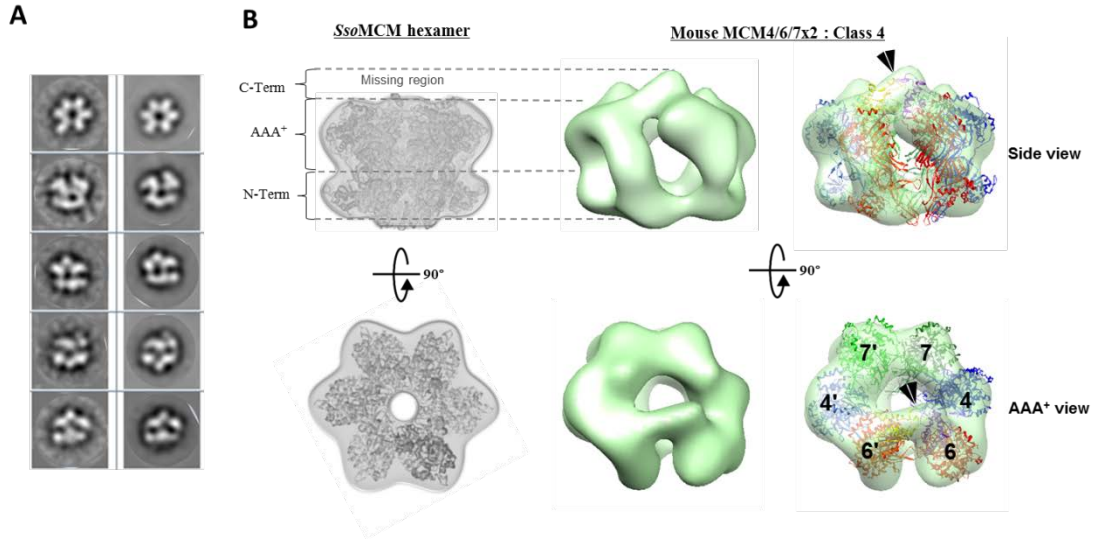


Figure 48. ***Class 4*****: A MCM4/6/7x2 loosely closed ring with a C-terminal connecting bridge. (A)** Class average of experimentally acquired images (*left* column) and matching forward projections (*right* column) of the reconstructed model. **(B)** From left to right there are shown the side views (*top* panels) and AAA$^+$ views (*bottom* panels) of: an *in silico* assembled hexamer model of *Sso*MCM lacking the C-terminal region (modified from (Brewster et al. 2008)); a solid colored 3D map of MCM4/6/7x2 model *Class 4*, and the fitting within the same map of six copies of *Sso*MCM monomer (PDB-3F9V, in orange, light-blue, light-green, red, blue and green). Black double-arrowhead points to an extra density that creates a connecting bridge between the putative AAA$^+$ regions from two adjacent subunits, where two copies of the HTH motif from *Hs*MCM6 (PDB-2KLQ, in light orange and purple) can be fitted. Numbers represent the putative organization of the different subunits.

Class 5: A closed ring with a N-terminal sealed pore and a big lump at its C-terminal region

In *Class 5* map (Figure 49), the opening of the central channel at the AAA$^+$ region of the ring is quite wide (diameter of around 50Å), contrasting with its N-terminal region, which is constricted in such a way that is practically sealing the pore at this side, a feature that will avoid DNA to cross along the helicase central pore. Another remarkable feature is a prominent mass placed at the C-terminal region, extruding between two subunits. Given that two copies of the HTH motif of *Hs*MCM6 (PDB-3F9V) can be accommodated within the lump, and following the previously exposed reasoning for map *Class 1*, we propose that this extra mass could be designated to be formed by the interacting C-terminal motifs of the putative MCM6'-6 protomers, positioned at

both sides of the mass. The model presented in Figure 10 B was followed to complete the arrangement of the remaining subunit.

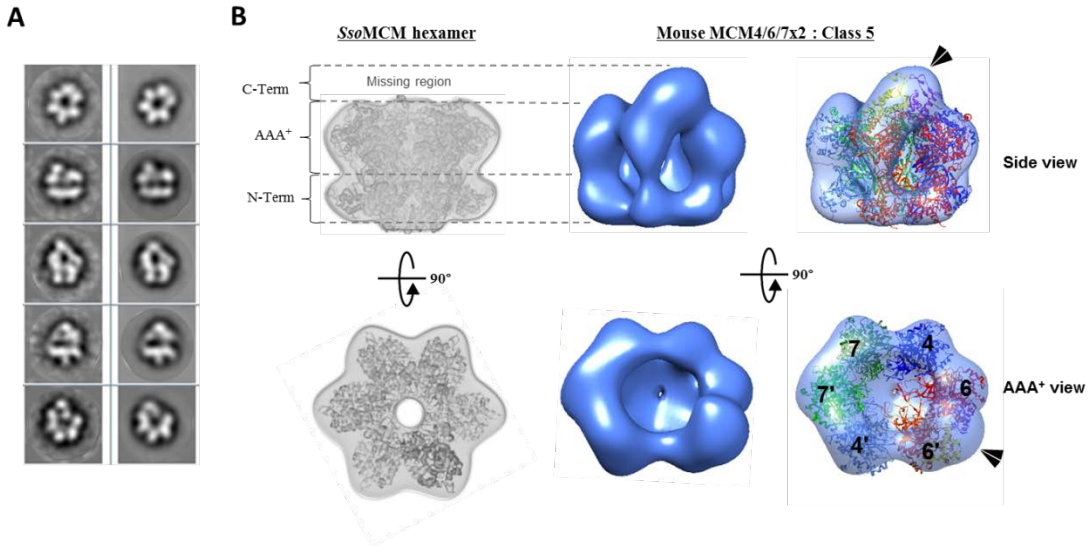


Figure 49. *Class 5*: N-terminal sealed MCM4/6/7x2 ring with a big C-Ternimal electron density. (A) Reference-free class averages (*left* column) and matching forward projections (*right* column) of the density map. **(B)** From left to right there are shown the side views (*top* panels) and AAA$^+$ views (*bottom* panels) of: an *in silico* assembled hexamer model of *Sso*MCM lacking the C-terminal region (modified from (Brewster et al. 2008)); a solid colored 3D map of MCM4/6/7x2 model *Class 5*, and the fitting within the same map of six copies of *Sso*MCM monomer (PDB-3F9V, in orange, light-blue, light-green, red, blue and green). Black double-arrowhead points to a big extra density at the putative C-terminal region of the complex, where two copies of the HTH motif from *Hs*MCM6 (PDB-2KLQ, in light orange and purple) can be fitted. Numbers represent the putative organization of the different subunits.

Class 6: A notched ring with a big lump at its C-terminal region

Class 6 map (Figure 50) shows the structure of a ring with a discontinuity at the AAA$^+$ region. One of the protomers that resides at the nick have an extra mass protruding at the C-terminus, in which two copies of the HTH structure of MCM6 (PDB-2KLQ) can be fitted, similarly to what was previously observed in the maps of *Class 1* and *Class 5*. *Class 6* EM map resembles the shape of the planar, notched conformation of a Drosophila MCM2-7 map (EMD-1834) when displayed at a lower contour level (Figure 50 C), where it is observed an extra mass departing from the MCM6 AAA$^+$ domain, although in this case it extends across the nick reaching a position over the next subunit and, as expected, allows to fit only one copy of the HTH structure (PDB-3F9V) (Costa et al. 2011). Considering all above observations and the model of the subunit distribution of MCM4/6/7x2 (Figure 7), we propose, again, that the extra mass corresponds to the interacting C-

terminals of both MCM6'-6 subunits and that the two subunits at the AAA$^+$ notch interface should be the putative MCM4'-6'. A nick or weak association between MCM4 and MCM6 has also been observed in EM models of *Dme*MCM2-7 (Costa et al. 2011) and yeast OCCM complex (Sun et al. 2013), which reflects the trend of these subunits to dissociate.

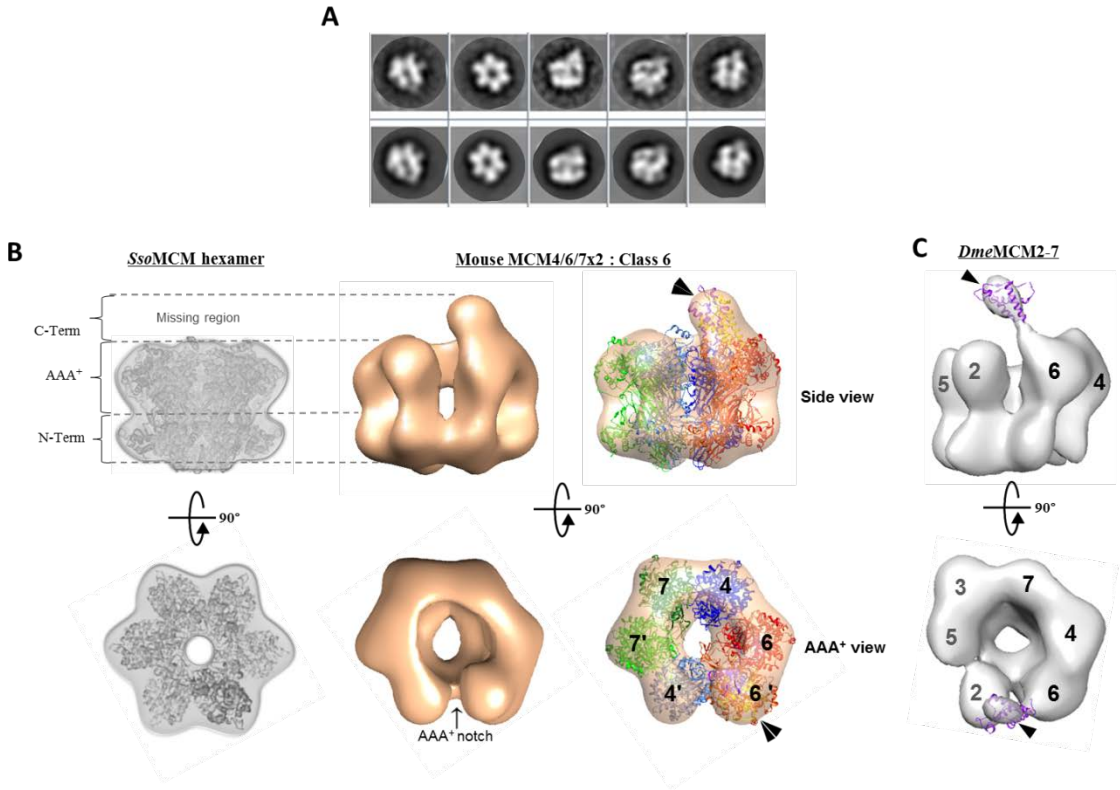


Figure 50. *Class 6*: A MCM4/6/7x2 complex with a big additional density protruding from one side of a notched AAA$^+$ region. (A) Reference-free class averages (first row) and the corresponding forward projections (second row) of the 3D map. **(B)** From left to right there are shown the side views (*top* panels) and AAA$^+$ views (*bottom* panels) of: an *in silico* assembled hexamer model of *Sso*MCM lacking the C-terminal region (modified from (Brewster et al. 2008)); a solid colored 3D map of MCM4/6/7x2 model *Class 6*, and the fitting within the same map of six copies of *Sso*MCM monomer (PDB-3F9V, in orange, light-blue, light-green, red, blue and green). Black double-arrowhead points to a big extra density at the putative C-terminal region of the complex, where two copies of the HTH motif from *Hs*MCM6 (PDB-2KLQ, in light orange and purple) can be fitted. **(C)** Structure of *apo Dme*MCM2-7 (EMD-1834). Black arrowhead points to a C-terminal extra density protruding from MCM6, where it is fitted the structure of the *Hs*MCM6 HTH motif (PDB-2KLQ. Purple). In all cases, numbers represent the different subunits.

4.3. Human Twinkle

4.3.1. Aggregation of helicase particles occurring at low salt concentration is reversible by increasing the ionic strength

The solubility of Twinkle is greatly affected by the ionic strength of the media (Ziebarth et al. 2010). Even in presence of cofactors ($MgCl_2$ and ATPγ-S) that help stabilize Twinkle, low concentrations of salt results in an important loss of soluble protein. On the other hand, increasing the temperature along a range of 4-37°C have shown to enhance the recovery of soluble protein when using concentrations above 150 mM NaCl (Ziebarth et al. 2010).

In order to have a direct and detailed visualization of the effect of different salt concentrations on the protein state, we monitored by EM a soluble sample at 330 mM NaCl (without adding cofactors), and then we diluted it until reaching 150 mM NaCl and 100 mM NaCl concentrations, adding $MgCl_2$ and ATPγ-S as stabilizing cofactors (Figure 51 A, B and C, respectively). NS-EM images of the sample at 330 mM NaCl (Figure 51 A) allowed a clear visualization of a heterogeneous group of free particles, most of which showed a star-like shape with a variable number of flexible arms (each monomer). The protomers are bound to each other by one end creating a central channel. A mixture of extended and more compacted particles was observed with no signs of sample aggregation. Conversely, NS-EM evaluation of the samples at 150 mM NaCl (Figure 51 B) and 100 nm NaCl (Figure 51 C) showed clear signs of an important degree of protein aggregation, which was more pronounced in the lower salt condition, corroborating that the level of aggregation is inversely proportional to the salt concentration of the sample (Ziebarth et al. 2010).

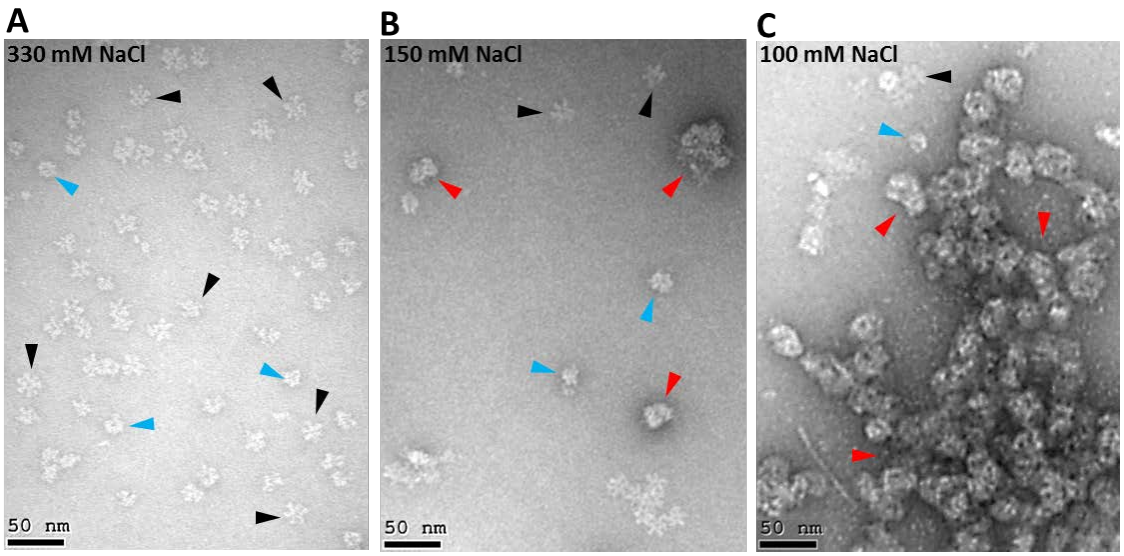


Figure 51. Representative NS-EM images showing the aggregation effect on a wild type Twinkle sample under three different concentrations of salt. (A) 330 mM NaCl, **(B)** 150 mM NaCl + Cofactors and **(C)** 100 mM NaCl + cofactors. Black and blue arrowheads point to either extended or more compacted helicase particles, respectively, while red arrowheads point to protein aggregates.

We were interested to know whether the protein aggregates observed at the lowest salt conditions could be solubilized back upon increasing salt concentration. For that purpose, we started from a partially aggregated sample at 100 mM NaCl plus cofactors (Figure 52 A), and we increased the salt concentration to 330 mM NaCl (Figure 52 B). Remarkably, it was observed not only that the aggregation disappeared, but also that particles showed their initial appearance of ring-shaped complexes with flexible arms.

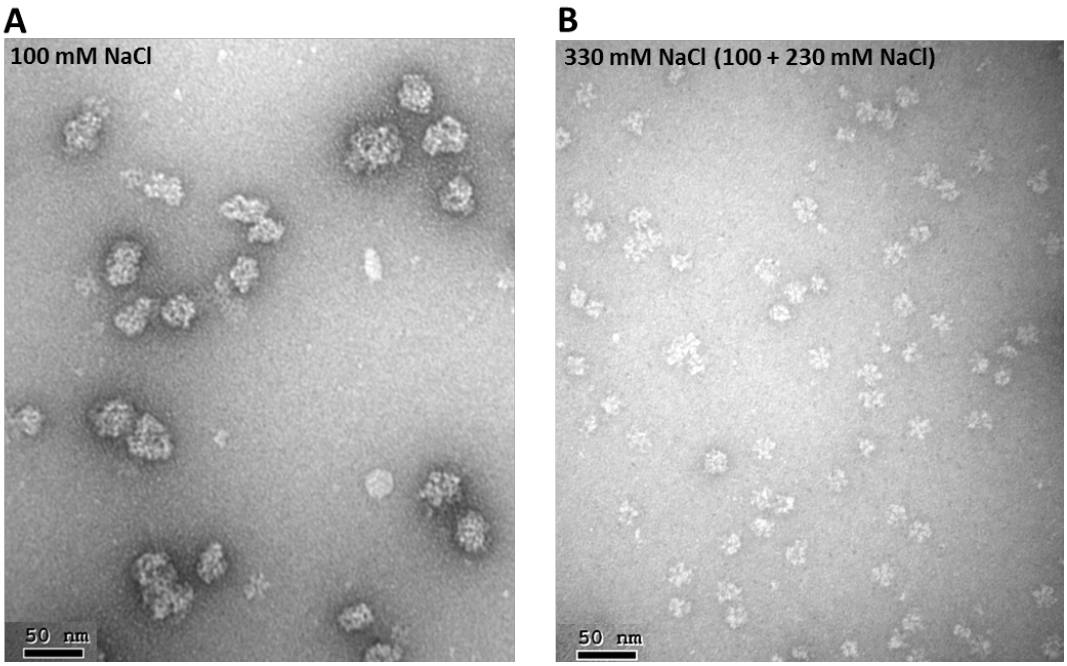


Figure 52. Recovery of helicase particles from an aggregated sample. Representative NS-EM micrographs showing **(A)** a protein sample almost completely aggregated at 100 mM NaCl + cofactors, and **(B)** the same sample after increasing the salt concentration to 330 mM NaCl, where it can be seen the complete recovery of soluble helicase particles.

4.3.2. Twinkle forms multiple homo-oligomeric complexes showing long and flexible arms

The 2D image analysis and classification of 14,500 single particles selected from the full-length Twinkle sample under 330mM NaCl conditions (Figure 53), revealed a heterogeneous collection of particles composed of different oligomers with star-like shapes showing both variable extended and more compacted conformations (Figure 55). The compact particles constitute around 35% of the total, while only around 15% of the particles have all the protomers completely extended; the remaining 50% was formed by particles with a mix of extended and flexed monomers.

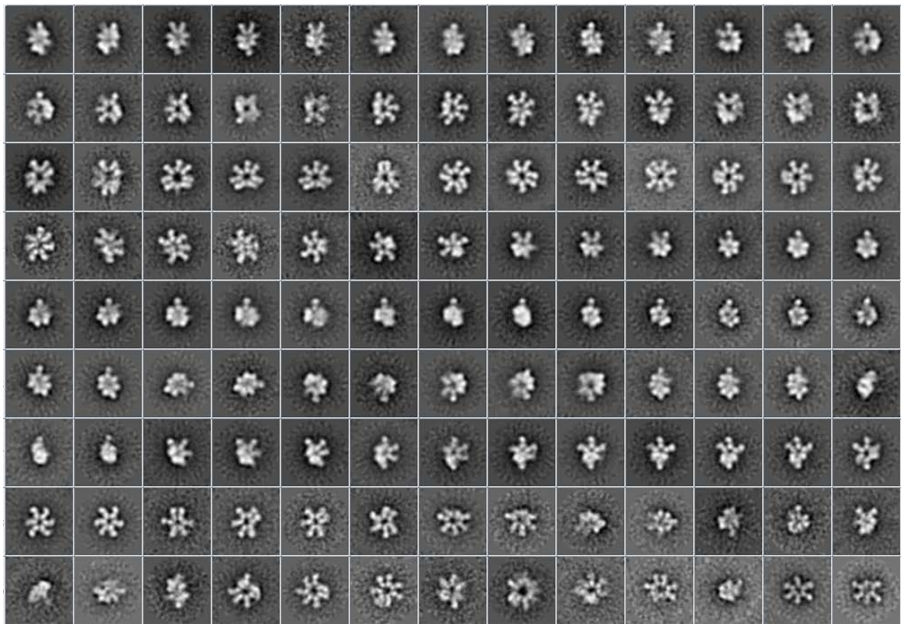


Figure 53. Reference-free class average classification (CL2D) of the complete set of 14500 particle images of the full-length Twinkle protein sample at 330 mM NaCl.

Most of the observed images were frontal views of the particles (Figure 53) and, based on longitudinal measurements and the classical two-tier shape of most ring-like helicases, we concluded that only a limited number of particles (about 5%) corresponded to side-views of compact conformations (Figure 54). Possible side-views of the extended particles were only observed in 55° tilted images, which was not surprising considering that the extended conformations seems to have a relatively flat disk-like shape.

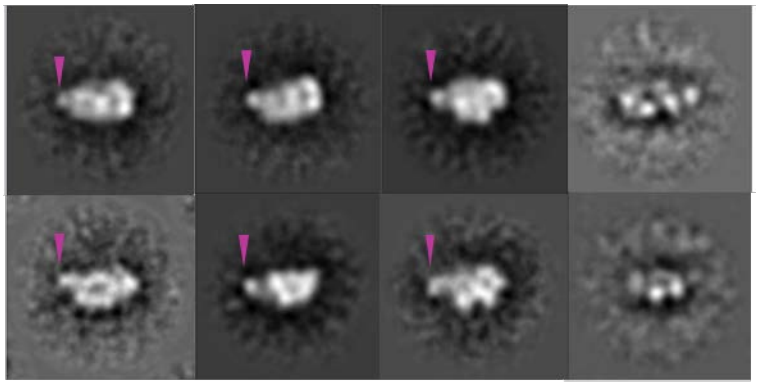


Figure 54. Lateral views of the wild type (WT) Twinkle. NS-EM 2D class averages showing possible lateral views of the particles, where a density that could correspond to the NTD of an extended protomer is pointed with a fuchsia arrowhead. The two images in the last column at the right side of the panel came from 55°tilted images.

In addition to the already reported hexameric and heptameric complexes formed by Twinkle (Ziebarth et al. 2010), (Fernández-Millán et al. 2015), here we identified a new octameric species form of the protein (Figure 55 A). Similarly to the other two lower order oligomers, the octamers presented a radial organization of its monomers, but it was a minority class with only 5% of the images.

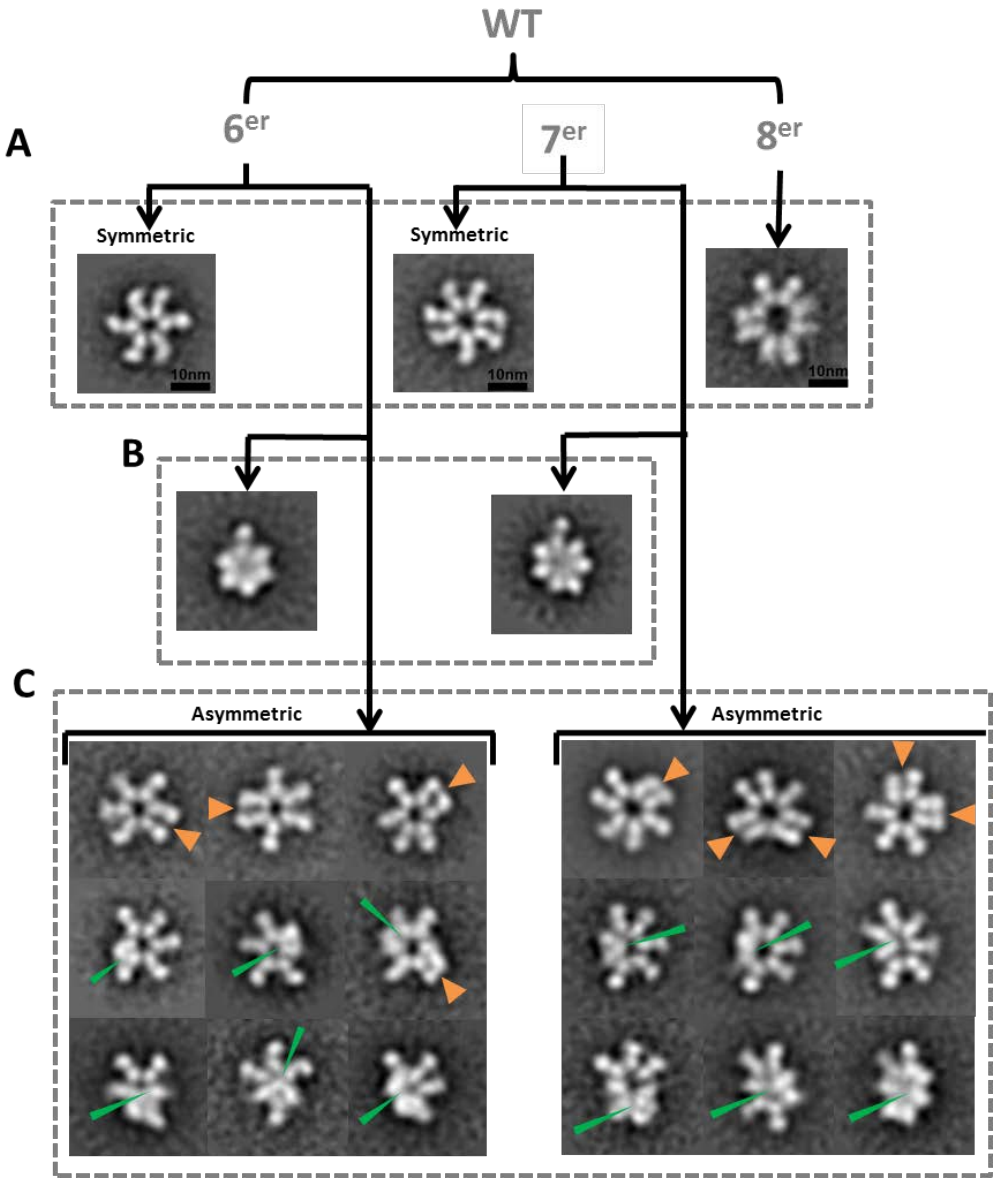


Figure 55. Diagram showing representative NS-EM 2D classes of the wild type (WT) Twinkle sample at 330 mM NaCl. (A) Different type of oligomers (hexamer, 6^{er}; heptamer, 7^{er} and octamer, 8^{er}) formed by the protein. **(B)** Compacted particles with only one arm extended. **(C)** Variable conformational states where extended protomers that tend to form "close pairs" are pointed by orange arrowheads. Some of the protomers seem to be bent toward the center of the complex (pointed by a green arrowhead).

The observed sample population, mostly formed by similar proportions of hexamers and heptamers (Table 3), was in agreement with the expected oligomeric forms of the selected purified fractions coming from the glycerol gradients.

Oligomer state	WT (%)
Hexamer	45
Heptamer	50
Octamer	5

Table 3. Approximate percentage (%) of the different oligomeric state particles for the full-length wild type (WT) Twinkle sample.

Both the crystal structure of a T7 gp4 heptamer (Toth et al. 2003) and the 3D map of a chemically fixed hexameric form of Twinkle (Fernández-Millán et al. 2015) show a compact, modular arrangement where the CTDs form a closed ring topped by the NTDs loosely arrayed (Figure 14). Contrasting with the above mentioned compact conformations, dynamic studies of Twinkle in solution have shown that this helicase tends to acquire an extended conformation where the flexible NTDs are primarily orientated on the sides of the CTD ring (Fernández-Millán et al. 2015), which is in agreement with our NS-EM images showing mostly extended particles of Twinkle. We observed that each monomer has an articulated arm-like structure with two main globular electron densities separated by a soft density in between, which may correspond to the flexible linker connecting the CTD and the NTD. Within a Twinkle helicase particle, each of its protomers binds to the next one by an edge lobe putatively corresponding to the oligomerization CTD, creating a central channel. The lobes at the opposite side of the central channel are putatively designate as the NTDs, and showed a wide number of different orientations. A simplified, general description of the domain´s arrangement within the complexes can be made based on the more symmetric, arms-extended conformations (Table 4). These extended particles have a central channel composed by the putatively closely interacting CTDs, followed by a final ring area occupied by the more flexible putative NTDs. The diameter of the central channel increased accordingly with the oligomeric state of the particles, and for the three cases, hexamer, heptamer and octamers, their pore is big enough to easily accommodate ssDNA or dsDNA.

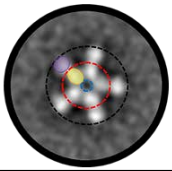
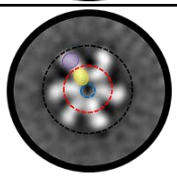
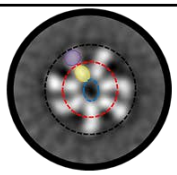
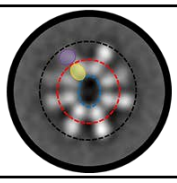
Oligomeric state	External diameter (nm)		
	Central channel	CTD ring	NTD ring
Pentamer 	~2	~11	~20
Hexamer 	~3	~12.5	~21
Heptamer 	~3 x 5	~13.5	~22
Octamer 	~7.5 x 5	~14.5	~24

Table 4. Colored dashed circles demark the approximate external diameter of the particles' central channel (blue), CDT (red) and NTD (black) rings, for each of the different oligomeric states. A yellow and a purple oval, respectively highlight the putative CTD and NDT of one monomer.

Though very subtle, some of the subunits show a small density positioned at one side of the NTD lobe. It is noteworthy that the NTD lobes presenting this extra mass, look less dense than most of the ones where only the larger globule is observed, which suggest that this part may be formed by two modules that tend to be very close but sometimes can be more separated (Figure 56). Considering that the NTD is structurally divided into a bulky RPD and a much smaller and flexible ZBD at the terminal edge (Fernández-Millán et al. 2015), the small density observed in our images could be the ZBD. The above is in agreement with the recently solved EM structure of a compact conformation of Twinkle (Fernández-Millán et al. 2015), where the detected ZBDs are positioned at one side of its respective RPD.

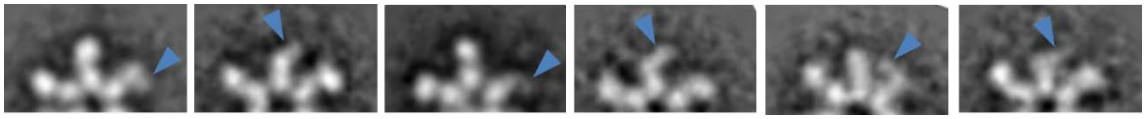


Figure 56. Small density at the NTD. Blue arrowheads point to a small extra density protruding at one side of the NTD lobe which, in turn, seems to be less electron dense that the NTDs that do not present the extra mass. By its position and size, the small extra mass could be the ZBD.

A closer analysis of the conformational variability in the full set of 2D class images suggests two main types of movement of the arms (its monomers) that can be combined in multiple ways. The first class of movement is an axial displacement of the NTDs, where there is an observed tendency of the extended arms to arrange in close pairs (Figure 55 C). The second form of movement is the bending of the arms, where a variable number of NTDs bent toward the ring area of the CTDs. Several combinations of bent NTDs occur, showing different number of both neighbouring and distant subunits contracted against the center of the complex (Figure 55 C). Whether each NTD is placed on its own RPD or the one of a neighbor subunit, is a question than cannot be unambiguously answered only with this 2D image analysis. However, the solved structure of both compact conformations of a Twinkle hexamer (Fernández-Millán et al. 2015) and a T7 gp4 heptamer (PDB-1Q57) (Toth et al. 2003), showed NTDs placed on top of the CTDs from a neighbouring protomer. A second question that arises is which of the interdomain interactions, ZBD/RPD or RPD/CTD, could be responsible for a determined type of observed movements, for which it could be useful a comparative analysis between the EM images of the wild type Twinkle and those coming from a protein lacking the ZBD.

It specially caught our attention the apparent absence of conformations with all the arms retracted. Instead, close to 20% of the particles showed a conformation where all but one (n-1) of the protomers bent toward the central pore letting only one arm extended with its NTD fully exposed to the solvent. The above described one-arm-extended conformation is in agreement with a conformation previously reported (Fernández-Millán et al. 2015), where a chemically fixed Twinkle sample show a complex with only one NTD isolated, while the remaining RPDs are interconnected through contacts with neighbouring ZDBs. The apparently structural predisposition of Twinkle to have one loosed monomer could be supported by the proposed ring-opening mechanism of T7 gp4 (Ahnert et al. 2000), where the initial binding of the DNA by the N-terminal

primase domain produces a conformational change that triggers the ring opening for its subsequent loading onto the DNA molecule, followed by the closing of the ring.

4.3.3. The ZBD has a role on the structural flexibility

It has been reported that deletion of ZBD of the human Twinkle have a negligible effect on the stability of the complex and its helicase activity (Farge et al. 2008), so the enzyme is still able to support nearly normal levels of DNA synthesis during *in vitro* conditions. On the other hand, the lack of ZBD reduces the helicase´s ssDNA binding ability, and consequently, its ssDNA-dependent stimulation for ATP hydrolysis (Farge et al. 2008).

We were interested in evaluating possible structural effects of ZBD beyond oligomerization. To this aim, we analyzed by NS-EM a Twinkle peptide lacking ZBD (ΔZBD, our protein construct Del.E147K684). The initial inspection of the sample showed similar images to those coming from the wild type protein (Figure 57), which is in agreement with the fact that ZBD is not required for the formation of higher order oligomers. No sign of sample aggregation was observed at 250 mM NaCl conditions.

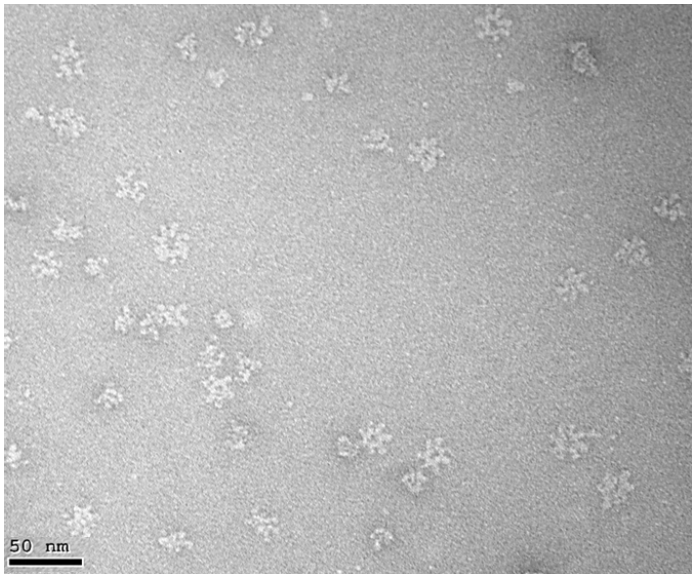


Figure 57. Representative NS-EM micrograph of the truncate ΔZDB protein at 250 mM NaCl.

9,991 single particles images of the negatively staining ΔZBD sample were subjected to reference-free 2D image classification (Figure 58).

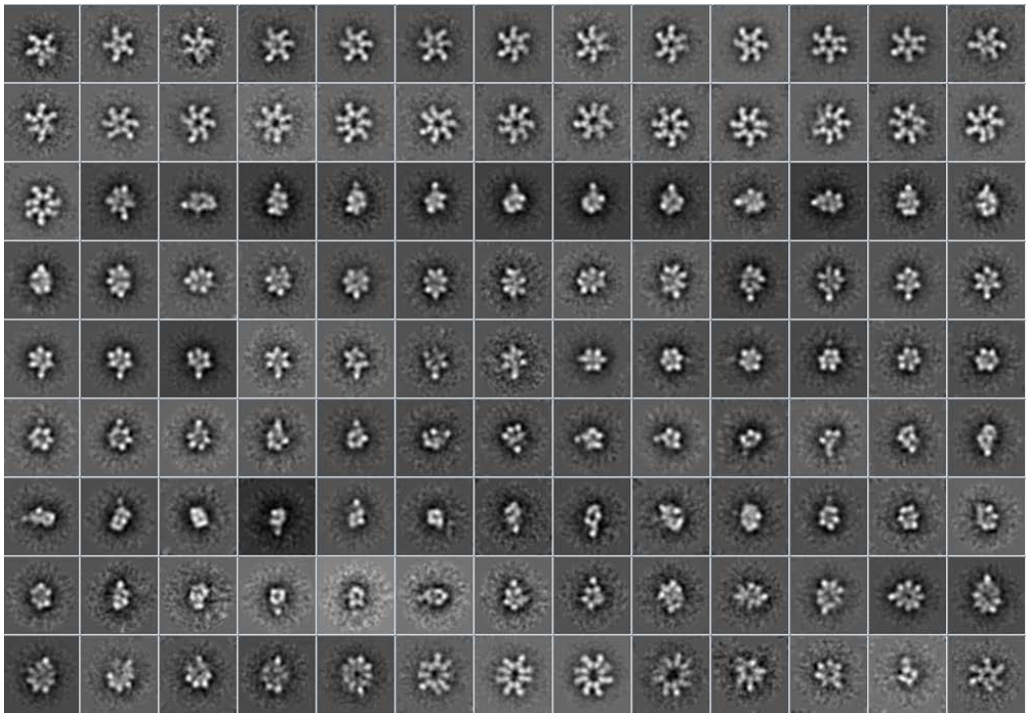


Figure 58. Reference-free 2D class average classification (CL2D) of the complete set of 9991 particle images of the Twinkle deletion construct lacking the ZBD (ΔZBD) of protein.

The analysis of the 2D classified images revealed important differences between the ΔZBD protein and the full-length Twinkle. Both completely extended and compacted particles are formed by the ΔZBD protein (Figure 59A), but aside from the complexes with only one arm extended, there were no other intermediate conformations with variable number of extended/flexed arms which, moreover, are very commonly formed by the full-length protein. Indeed, compared with the full-length protein, the extended conformations of the ΔZBD protein showed a relatively symmetric distribution of the arms where the NTDs are well separated one from each other. This contrasts with the mostly asymmetric extended particles observed with the full-length protein, where pairs of protomers showing NTDs in close proximity occur frequently.

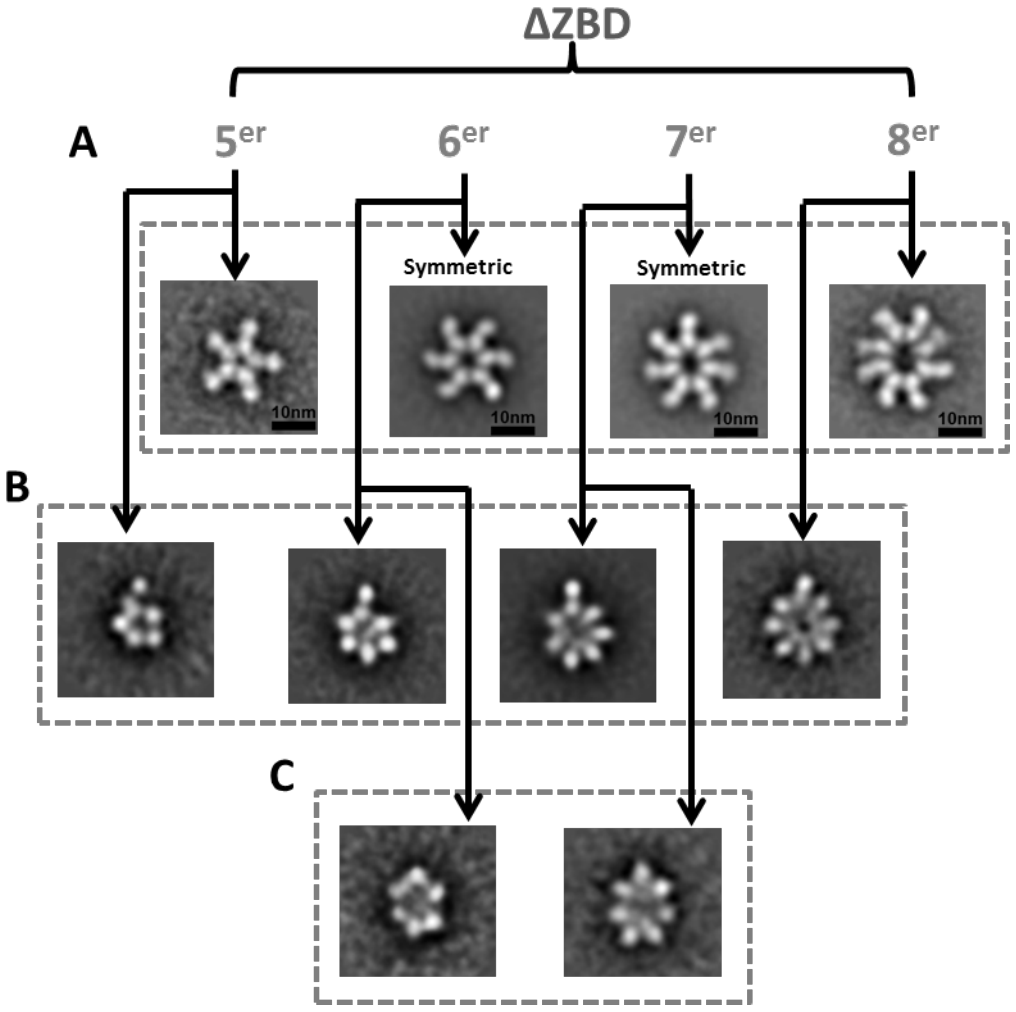


Figure 59. Diagram showing representative NS-EM 2D classes of the ΔZBD construct at 250 mM NaCl. (A) Different type of oligomers (pentamer, 5^{er}; hexamer, 6^{er}; heptamer, 7^{er} and octamer, 8^{er}) formed by the protein. **(B)** Compacted particles with only one arm extended. **(C)** Completely compacted particles. With all the arms flexed.

Additionally to the n-1 arms retracted conformation (Figure 59 B), the ΔZBD protein acquires conformations where all the monomers bend towards the center of the ring (Figure 59 C), which was not observed in the sample of the wild type Twinkle. While the number of hexamers of ΔZBD (Table 5) remained similar to that observed in the wild-type sample (Table 3), the amount of heptamers decreased considerably. Besides detecting around 10% of octameric particles, the ΔZBD protein also showed a similar percentage of pentamers (Figure 59 A and Table 5). Despite pentamers not being detected in our wild type sample under the used conditions, Twinkle pentamers have been previously reported (Ziebarth et al. 2010) by SDS-PAGE analysis under conditions of 100 mM NaCl plus cofactors ($MgCl_2$ and ATPγ-S).

Oligomer state	ΔZBD (%)
Pentamer	10
Hexamer	50
Heptamer	30
Octamer	10

Table 5. Approximate percentage (%) of the different oligomeric state particles for the ΔZBD construct sample.

We also observed a marked change in the proportion of extended vs compact conformations, being the compact ones more that 70% of the total of particles in the ΔZBD sample, while only 35% in the sample of the full-length protein. Concomitantly with an increase in the number of compact particles, it was observed a higher number of side-views (Figure 60) when compared with the wild type Twinkle, although it was still quite low to make a 3D reconstruction.

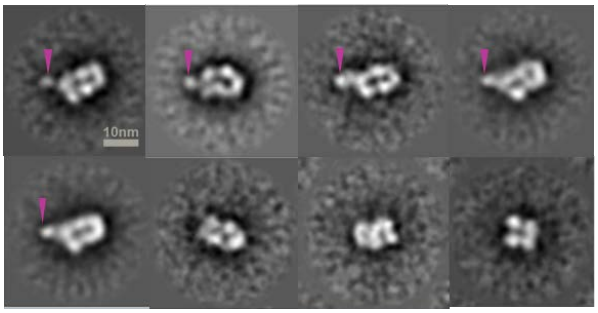


Figure 60. Lateral views of the ΔZBD construct. NS-EM 2D class averages showing possible lateral views of the particles, where a density that could correspond to the NTD of an extended protomer is pointed with a fuchsia arrowhead.

The observed conformational differences between the full-length and the ΔZBD proteins suggest that ZBD is implicated in a broad number of different conformations that do not occur in its absence.

4.3.4. Initial 3D maps: One and two floor structures

In agreement with previous reports (Ziebarth et al. 2010), (Fernández-Millán et al. 2015), our results show that Twinkle is a highly heterogeneous sample. The presence of multiple oligomeric and conformational states together with the quite low number of side-views of the particles, made a 3D reconstruction by NS-EM of the complexes unfeasible. Although the principal goal of the present work was making a 2D image comparative analysis between full-length and ΔZBD proteins, we were still interested in obtaining preliminary NS-EM 3D maps of the most representative conformations of Twinkle, in order to make further refinements using cryo-EM data. To overcome the problem of the reduced number of side views, tilted pair images of the negatively stained wild-type specimen were taken. 9,721 tilted pair particles were selected and subjected to 2D average classification. Representative 2D classes from hexameric and heptameric complexes from both fully extended conformations and compacted forms with only one extended arm were selected and, for each case, an initial 3D map was reconstructed by the RCT method (Figure 61). Consistently with the mass of hexameric and heptameric oligomers of Twinkle, the maps of the extended conformations showed a flat disc-like shape, while the contracted conformations presented bulky lateral views suggesting a two-floor structure with one thinner density projected outside, which may correspond to the extended protomer.

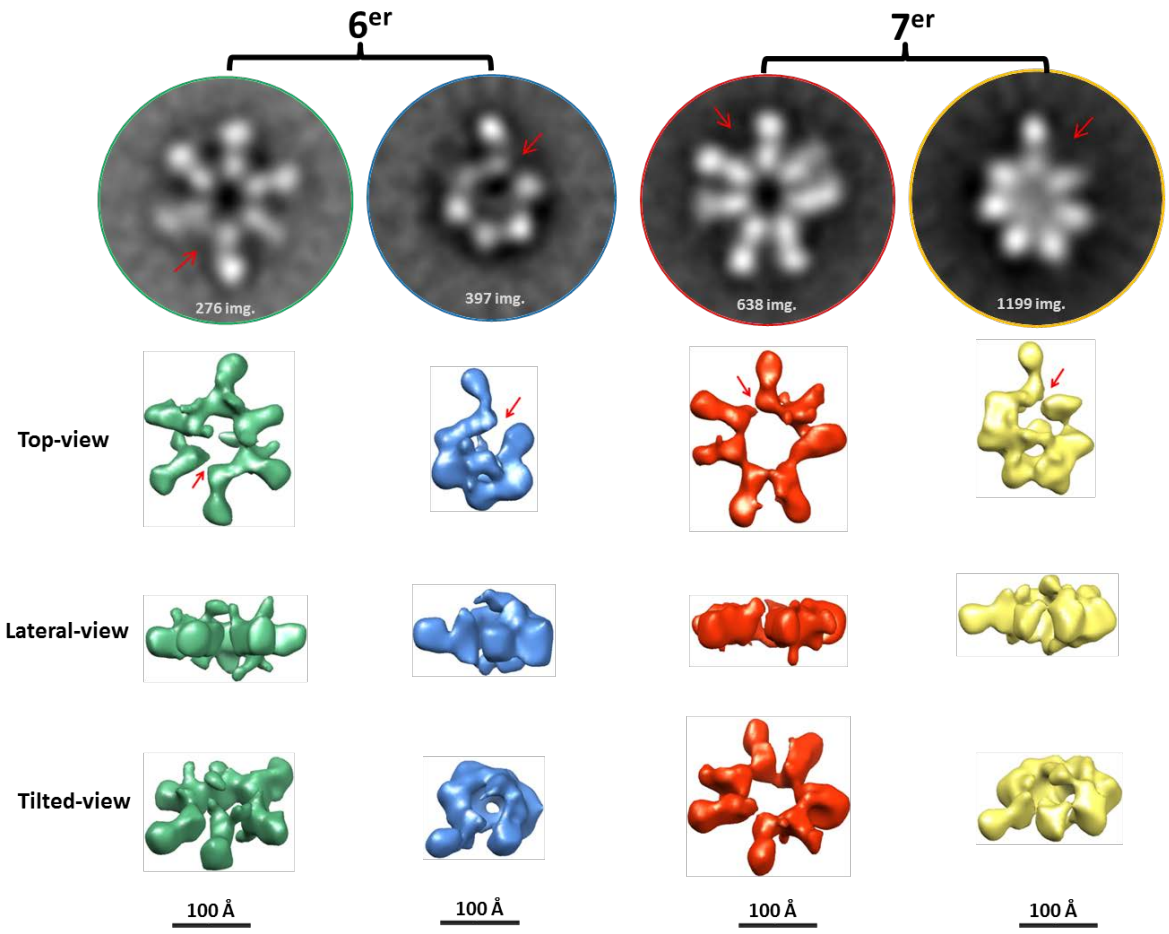


Figure 61. RCT reconstructions of both extended and compact conformations of hexameric (6^{er}) and heptameric (7^{er}) complexes. The first row showed the 2D classes whose corresponding 55° tilted pair images were used for obtaining the respective volumes (In white is showed the number of images on each class). A red arrow points to an area of lower density in the 3D maps (and the corresponding position in their 2D classes) which could suggests the presence of a thin gap.

All four 3D maps, especially those from hexamers, showed, apparently, one gap at the interface between two subunits. Although it is acknowledged that at the maps resolution leve,l this kind of detail should not be blindly trusted, in support of a possible ring opening some of our 2D average classes showed a thin gap (Figure 62) and, it has also been previously reported an open conformation of a Twinkle heptamer (Fernández-Millán et al. 2015). Indeed, there is a natural ability in other ring-shaped helicases to acquire open-ring conformations (Costa et al. 2011), (Arias-Palomo et al. 2013) as a mechanism for DNA loading, providing some level of functional support to the existence of a real opening or, at least, a weaker point of contact between two protomers in our Twinkle maps.

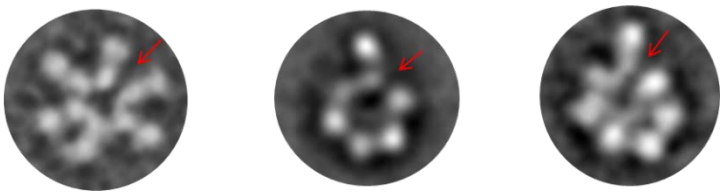


Figure 62. 2D classes of particles with an apparent thin gap (pointed by a red arrow) on the ring.

4.3.5. Establishing sample conditions for cryo-EM studies

Compared with NS-EM, cryo-EM is a higher resolution technique that avoids preferred orientations of the particles on the grid, so we started working to establish the cryo-EM conditions for future structural studies of Twinkle. Besides being highly sensitive to low salt concentrations, we also observed that in the absence of glycerol Twinkle forms a kind of inclusion bodies (Figure 63 A) that, interestingly, present some form of internal organization (Figure 63B), suggesting that the helicase complexes could still be assembled.

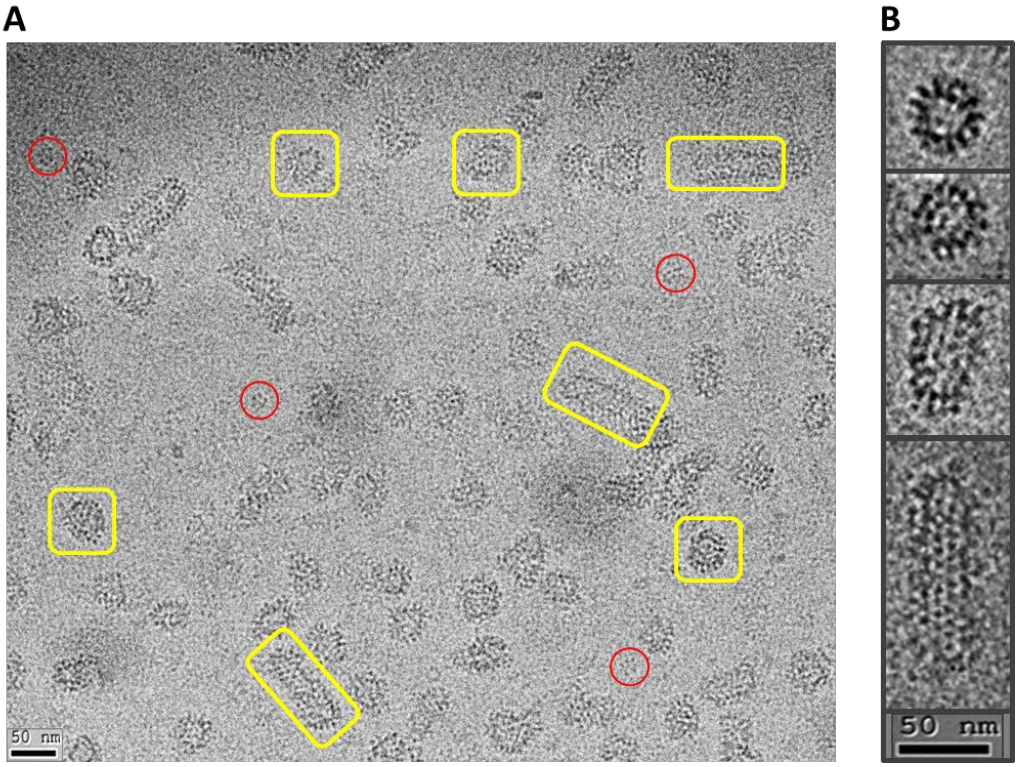


Figure 63. Representative cryo-EM images of Twinkle at 330mM NaCl and in absence of glycerol. (A) Micrograph showing protein aggregates (highlighted in yellow) of relatively regular rounded or elongated shapes. Some few free helicase particles can be identified (highlighted in red). **(B)** 2 x zoom of some aggregates where there can be inferred some form of internal organization.

Use of sample additives such as glycerol could be very problematic for vitrification due to its cryoprotectant effect. An additional problem of glycerol is the negative effect for cryo-EM imaging due to the contrast reduction it causes (Cong & Ludtke 2010). Because our initial sample came from a glycerol gradient purification, we had to work on finding the optimal equilibrium condition to maintain both the stability of the helicase complexes and an acceptable amount of glycerol for making cryo-EM grids. The final buffer conditions for cryo-EM were: 35 mM Tris-HCl, pH 7.5, 2 mM β-mercaptoethanol, ~5% glycerol and 330 mM NaCl, with a protein concentration of 50ng/µl. At this moment, we are in the process of cryo-EM image acquisition (Figure 64), so higher resolution maps for different oligomeric and conformational states of Twinkle are expected to be obtained in the near future.

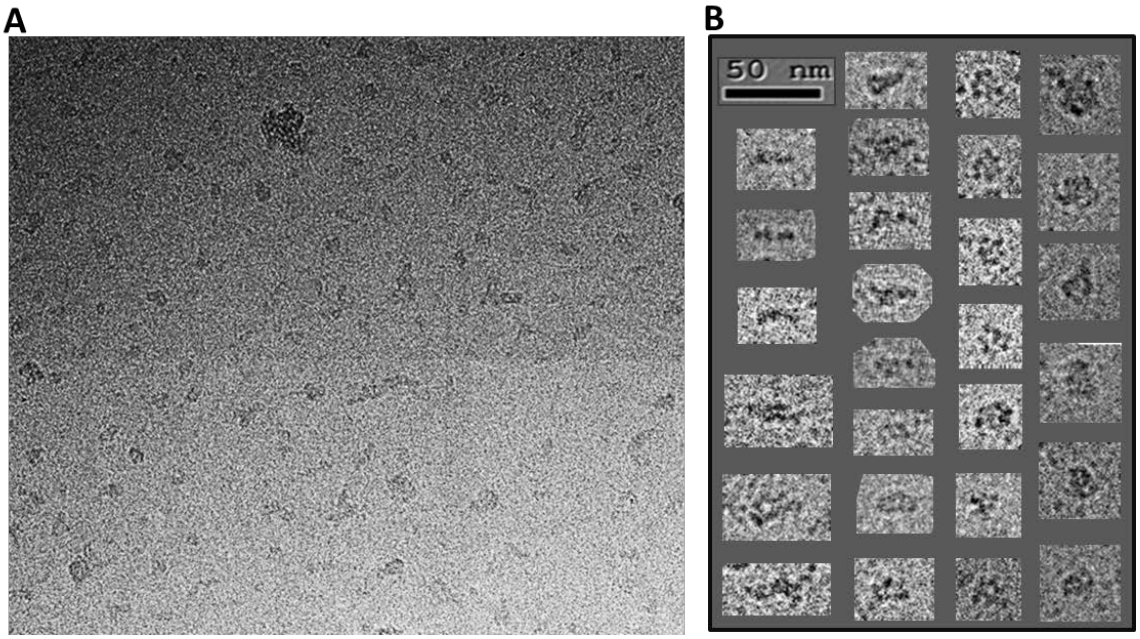


Figure 64. Cryo-EM of the wild type Twinkle sample at 330 mM NaCl. (A) Representative micrograph image. **(B)** Individual particles (zoomed images), where some relatively flat particles (first, left column) likely correspond to lateral-views of extended conformations.

Chapter 5

5. Discussion

5.1. Intrinsic protein flexibility facilitates quaternary structure assembly and conformational/oligomeric fluctuations

The assembly of proteins into higher order complexes is ubiquitous within the cell (Havugimana et al. 2012), (Robinson et al. 2007) and is required to perform many of the diverse biochemical activities essential to cell homeostasis, growth and proliferation. The intrinsic flexibility of proteins is intimately related to their functions, which often go hand in hand with their ability to interact with other molecules of the same (homomeric) or different (heteromeric) class to form either transient or stable complexes (Marsh & Teichmann 2014), (Marsh & Teichmann 2013). Furthermore, the flexibility of the unbound state of a protein generally correlates with the magnitude of binding-induced conformational changes (Marsh & Teichmann 2011), (Dobbins et al. 2008). In many cases, the subunits (that is, the individual polypeptide) constituents of a protein complex can be assembled into a wide variety of symmetric and/or asymmetric quaternary structure topologies, where each of them often carried out a different function. Although is expected that proteins tend to acquire a more stable and defined structure upon the interactions occurring within a macromolecular assembly, this does not imply the absolute loss of flexibility. Indeed, it is a common phenomenon that specific regions of a protein maintain a high degree of flexibilty in the new formed complex, and even in some cases, there are parts that remain unstructured, waiting to establish additional interaction with other molecules.

The three cases of study presented in this thesis; CPAP, MCM4/6/7 and Twinkle, constitute excellent examples of the natural structural versatility of proteins, showing the ability of establishing both flexible and multiple oligomeric complexes, which can be intuitively associated with the functional complexity of the biological processes where each of these macromolecular machines has been related with.

5.2. Structural insights into CPAP conformational and oligomeric behavior: The possible hierarchical building of a scaffold

The centrosome, a conserved organelle found in nearly every animal cell that acts as the primary microtubule organizing center (MTOC), is composed by a pair of centrioles immersed in a proteinaceous network of pericentriolar material (PCM). Centrosomes are involved in the organization of the mitotic spindle and in cytokinesis, as well as in the nucleation of cilia and flagella; they also serve as anchorage site for a wide number of regulatory processes.

CPAP (Centrosomal P4.1 associated protein, also known as Centromere Protein J –CENPJ-), a cell cycle regulated protein fundamental for centriole assembly, is a centrosomal protein present across the eukaryotic taxon. This protein is formed by one conserved globular domain at its C-teminus while the rest of the sequence is predicted to have disordered and coiled-coil regions. In humans, the carboxyl region of CPAP between the residues 897-1338 ($CPAP^{897-1338}$) is responsible for its interaction with several other important centrosomal proteins, and it includes a homodimerization domain. Mutations of CPAP have been associated with primary autosomal recessive microcephaly (MCPH). Despite of the widely demonstrated biological and clinical relevance of CPAP, its mechanistic behavior remains blurry.

Our biochemical, biophysical and electron microscopy work constitutes the first report of several higher-order oligomers of $CPAP^{897-1338}$ and allows us to provide new structural information of these complexes. Here we have shown and analyzed the 2D images of such structures and resolved the 3D-volume of both a putative $CPAP^{897-1338}$ dimer and a tetramer. Based in our observations, we propose that CPAP presents a discrete but dynamic oligomeric behavior, and that a tetrameric complex of the protein could be the structural brick of higher order supramolecular structures which could be working as a scaffold that tethers the PCM to centrioles.

5.2.1. CPAP$^{897\text{-}1338}$ forms different homo-holigomeric complexes *in vitro*

In this study we report by biochemical, biophysical and EM analysis that the human CPAP$^{897\text{-}1338}$ construct forms different types of homo-oligomeric complexes. Taking together the analysis of the data given by the SEC purifications; the 2D images and the 3D volume obtained by the EM work; the solved crystallographic structure of the G-Box domain; and, finally, the *in silico* structure modeling of the CC4/CC5 coiled-coils and a gross estimation of the dimensions of the CCGb-linker of CPAP, we propose that CPAP presents a discrete but dynamic oligomeric behavior that includes the formation of dimers, tetramers and larger structures formed by stacks of tetramers.

Taking into account the multiple binding partners of CPAP, it is not surprising that this protein is predicted to be largely unstructured, being this a characteristic that can confer the structural flexibility necessary to interact with different proteins/complexes. Indeed, it is reasonable to consider that each of the oligomeric states of CPAP would be more prone to establish an interaction with a specific protein/complex depending on its own initial configuration; the building up of different binding sites at the intersubunit interfaces of the different CPAP complexes could be a possible mechanism contributing to this.

It is well known that concentration can be a driving factor affecting the oligomerization status of many proteins, which in turn, modifies its basic structure and by definition also its function (Giese & Vierling 2002; Kley et al. 2011; Chen et al. 2003; Kutter et al. 2007). There exist several reports of proteins forming filaments in response to factors such as the increment in concentration (Noree et al. 2010) (molecular crowding), in some cases, showing a modular behavior (Petrovska et al. 2014) resembling the one we observe in CPAP$^{897\text{-}1338}$. Indeed, CPAP concentration is regulated along the cell cycle and its levels increase gradually from the beginning of S phase until mitosis, being this a lapse of time that coincides with the procentriole formation (in early S phase), elongation (in late S phase) and with centrosome maturation (along the G2 phase). Remarkably, centrosomal CPAP maintains a continuous exchange with a cytoplasmic CPAP pool (Kitagawa et al. 2011), reaching its highest level in G2, when there is a maximal recruitment of proteins to the PCM. CPAP shows a significant decrease at the end of mitosis/early G1, when the dynamic of formation and maturation of the centrosome is already finished (Tang et al. 2009), (Azimzadeh &

Bornens 2007), (Kim et al. 2012). Furthermore, the CPAP protein level is also regulated during centriole amplification in multiciliated cells (Zhao et al. 2013). Protein concentration-dependent conformational changes could be a mechanism contributing to direct the network of diverse interactions that CPAP must carry out. Thus, is tempting to speculate that cell cycle regulation of CPAP concentration could be a complex and highly synchronized strategy, which together with other determinant factors (e.g., protein phosphorylation (Chang et al. 2010), (Chen et al. 2006), (Zhao et al. 2010)) controls the oligomeric state of CPAP in order to direct some of its multiple functions.

Column-like structures with a repeating unit around 8nm have been observed in the inner walls of MTs in a tomographic reconstruction of the basal body triplet in *C. reinhardtii* (Li et al. 2012), and it has been suggested that these elongated structures could correspond to CPAP (Hatzopoulos et al. 2013). Furthermore, we note that tomographic images of centrioles from calf thymus (Figure 65 A), purified in our group, show modular, column-like structures whit an axial periodicity of approximately 80 Å thymus (Figure 65 C), running along the inner walls of MTs. These long centriolar structures look very similar to our purified stacks of $CPAP^{897-1338}$ (Figure 65 B) and, indeed, their repeated unit of close to 80 Å, match the dimension of the narrow side of the putative $CPAP^{897-1338}$ tetramers.

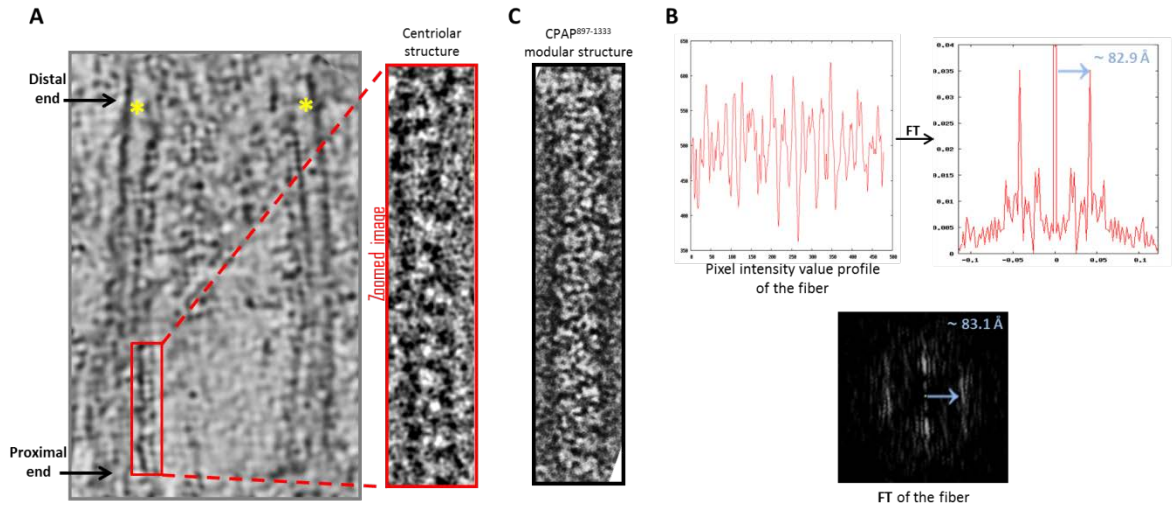


Figure 65. Centriolar localization of long, modular structures, proposed to be putative stacks of CPAP tetramers. (A) Highly filtered image of a cross-section of a calf thymus´ centriole thomogram (*left panel*), where part of a modular column-like structure localized at the inner wall of the centriole is highlighted in red. The proximal end of the centriole is at the bottom and the distal end is at the top. The position of the centiole walls are marked with a yellow asterisk (*). Zoomed image of the column-like structure (red panel), which would be putatively formed by stacks of CPAP tetramers. **(B)** NS-EM image of a long, modular structures formed by the purified $CPAP^{897-1338}$ protein. This mage is displayed at a similar scale to the one of the centriolar zoomed image showed in subparagraph **A**. **(C)** The centriolar modular structure has a ~ 83 Å axial repetition pattern (*left pannel*), as measured by both the Fast Fourier Transform (FFT) of the pixel intensity profile of the image (*top panel*) and by the diffraction pattern of the structure in the Fourier space (Fourier transform, FT) (*bottom panel*). The magnitude of the pattern is represented with a blue arrow and the value is written in the same color.

CPAP interacts with the 8 nm length α/β-tubulin heterodimers (Hung et al. 2004) that form the MT that constitute the walls of the centriole. It could be expected that proteins/complexes that interact with MT have patterns related to the one of the MT themselves. We propose that the interaction of putative CPAP tetramer stacks with the α/β-tubulin heterodimer repeats occurring along centriolar MTs every ~8 nm, would fix the structure of the modular CPAP filaments producing the 83 Å axial periodicity pattern observed in the putative CPAP column-like structure localized at the inner walls of the calf thymus´ centriole (Figure 65 A). Studies showing that CPAP concentrate within the proximal lumen of both parental and nascent centrioles (Kleylein-sohn et al. 2007),(Chang et al. 2010) support our proposed localization of the CPAP modular fibers, although the identification of CPAP within the purified calf thymus´ centrioles by direct labeling of the protein must be done.

An overview of our results and its analysis allowed us to propose a putative model for the formation of higher order modular filaments of CPAP (Figure 66), where first, the highly flexible

monomer strings of CPAP$^{897-1338}$ dimerize acquiring a globular and more stable toroidal structure. Then, dimers of CPAP$^{897-1338}$ would dimerize into a tetrameric structure, which, in turn, could acts as a somewhat flexible building block for larger, modular and elongated rope-like supramolecular structures with an axial periodicity around 8 nm. Finally, we propose that these modular rope-like structures formed *in vitro* could correspond to the column-like structures observed in the inner-walls of the centriole.

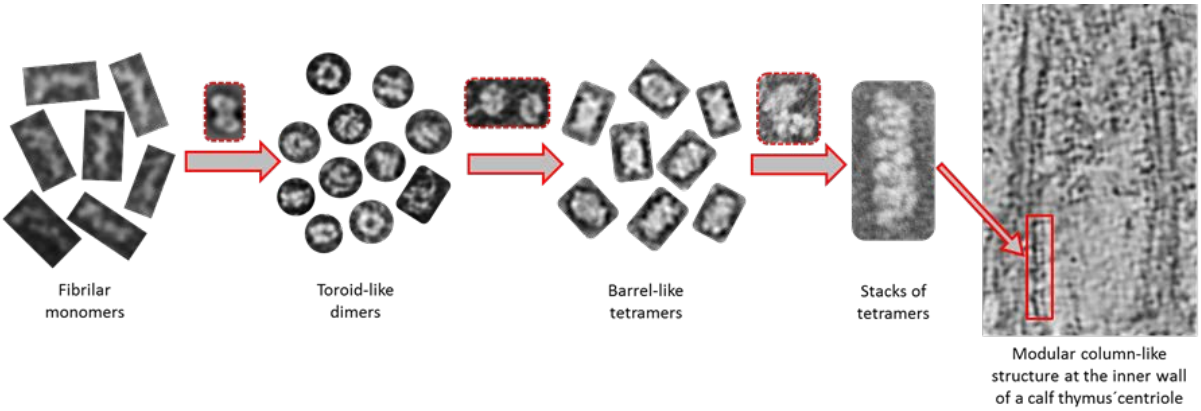


Figure 66. Tentative model for the progressive self-assembly of CPAP into gradually higher oligomeric subcomplexes until formation of a modular rope-like structure. The highly flexible monomers may dimerize by meshing into a globular toroid-like complex, which in turn could dimerize again resulting in the formation of a hollow, barrel-like tetrameric structure. Finally, stacks of tetramers may form linear arrangements of variable length with a periodicity close to 80 Å, which could also correspond to those structures known to exist at the inner walls of the centriole. Red dashed lines highlight putative intermediate states between one oligomeric state and the next one.

The proposed tentative organization of the G-Box and CC4/CC5 domains on opposite sides of the putative tetramer structure of CPAP$^{897-1338}$ would confer to it a structural (and functional) bipolarity, which, in turn, would also apply to the modular, higher order rope-like complexes. However, further work on CPAP´s domains labeling and the respective structural studies must be carried out to unambiguously determine the organization of the monomers within the different complexes.

The presented work reinforces the idea that CPAP forms organized higher order structures allowing it to act as an scaffold that connects PCM proteins and complexes with the nascent centriole (Gopalakrishnan et al. 2011; Hatzopoulos et al. 2013; X. Zheng et al. 2014), contributing to the progression of procentriole assembly and elongation (Tang et al. 2009). We propose that molecular crowding (Kuznetsova et al. 2014), (Ellis & Ellis 2001) could be an important driving force affecting both the structure and oligomeric state of CPAP. The precise mechanism by which

CPAP goes from one oligomeric state to a different one, as well as the exact biological role of each of these complexes, is a matter of future studies that would include works with CPAP mutants as well as interaction assays with some of its already known partners. Our results showing that $CPAP^{897-1338}$ displays a diverse and dynamic oligomeric behavior lay the foundation not only for future structural works, but also for new mechanistic/functional studies of this critical factor in centrosome formation and, consequently, in cell biogenesis. Mutations of CPAP, which could potentially affect its structure and ability to associate with other proteins, are associated with primary autosomal recessive microcephaly disease, thus the present study may lead to important future biomedical implications.

5.3. Insights into the architecture and flexibility of the hexameric MCM4/6/7 ring and other lower order sub complexes

The minichromosome maintenance (MCM) family is a conserved group of proteins proposed to work as replicative helicases in archaea and eukaryotes. They are usually assembled as ring-shaped molecular motors that are essential in the initiation and progression steps of DNA replication. In eukaryotes, each of the six homologous proteins MCM2-7 present different catalytic and regulatory roles and all of them are necessary to maintain the integrity of the genome. Eukaryote MCMs can form different types of oligomers, though it is not clear the biological role of most of them. A growing number of studies suggest that beyond replication, MCMs proteins must also be involved in DNA transcription, chromatin remodeling, activation of dormant replication origins and checkpoint responses to maintain genome stability. The hexameric MCM4/6/7 is a subassembly formed by two MCM4/6/7 heterotrimers that has an intrinsic helicase activity. There are many studies focused on the biochemical properties of the MCM4/6/7 hexameric helicase; however, structural works are practically inexistent. Here we present a set of electron microscopy (EM) maps that include different conformational states of the mouse hexameric MCM4/6/7 helicase, as well as some putative trimeric and monomeric structures. In agreement with previous biochemical studies of MCM4/6/7, and based on feature comparisons between our volumes and equivalent proteins in several of the density maps already resolved for the MCM2-7 heteroexameric helicase, we propose a putative architecture of MCM4/6/7 hexamer as well as the possible identification of the two protomers flanking a gap present in the MCM4/6/7 helicase. Finally, we present a tentative time sequence of the structures reported in this work, which attempt to explain a process beginning with the dimerization of trimers and ending with the formation of a closed, planar ring.

5.3.1. Structural polymorphism

A number of EM maps have shown the helicase MCM2-7 as a single or double hexamer displaying conformational variations, which can be also found in complex with other proteins of DNA molecules. On the other hand, there is very scant structural information on the helicase

MCM4/6/7x2, which is restricted to some studies presenting some EM 2D images of this complex (Bochman & Schwacha 2007), (Sato et al. 2000), (Yabuta et al. 2003). The fact that the trimer MCM7→4→6 forms part of the MCM2-7 helicase (Figure 10 A) lets us think that the already reported 3D maps of this complex can provide important clues about the structure and subunit distribution within MCM4/6/7x2. Furthermore, proteins MCM4, 6 and 7 can associate in to hexamers and trimers; even some of them have been purified as free monomers (Musahl et al. 1995), (Sherman et al. 1998), (Ma et al. 2010), (Ichinose 1996), (Bochman et al. 2008), (Davey et al. 2003), (Xu et al. 2013), but currently there are no solved structures for any of these states.

Up to now the structural observation of the C- and N-terminal extension of full-length eukaryotic MCM proteins has been elusive due to their high flexibility. The identification of these extensions within the context of the whole MCM4/6/7x2 helicase complex can provide unique important clues about its functional role, hence the relevance of working on this front.

It has been observed in different helicases that binding of nucleotide can induce the formation of active toroidal oligomers which, along with the hydrolysis process, brings about additional conformational changes required to carry out the enzyme functions (Hydrolysis et al. 1984), (Bacteriophage et al. 1995), (Ziebarth et al. 2010). The mouse MCM4/6/7 sample analyzed during this work, which was initially purified as an hexamer in the presence of ATP (which by the moment of our EM analysis might be potentially hydrolyzed), allowed us to visualize different structures of the hexameric ring as well as smaller oligomeric complexes that include putative trimers and monomers.

In this work, aside from presenting mouse MCM4, 6 and 7 low resolution 3D maps of a number of trimeric and hexameric complexes, we have carefully made an exercise of putting them in the context of the available structural and functional works on MCM proteins, which allow us to putatively locate the relative position among subunits and speculate about some functional consequences of the observed structural features. The presented volumes are EM snapshots that suggest transition states going from the isolated trimers MCM7→4→6 and MCM6'→4'→7' to the formation of a planar ring with a MCM4 C-terminal domain extended toward the central channel, and the interacting HTH motifs of the subunits MCM6 and 6' protruding, maintaining weak contact with the AAA^+ region of MCM4'. We speculate that this structure may be able to isomerize to an open ring with a MCM4'/6'gap, which could be a suitable conformation for DNA-loading. Due to

the putative architecture similarities of this MCM4/6/7x2 volume with MCM2-7, we propose that the subunit rearrangement within this ring conformation could be close to the active state required by the helicase to bind and, subsequently, unwind dsDNA, although it will be expected that upon DNA binding additional conformational changes take place. For example, in *Dme*CMG it is observed that the AAA^+ region of the MCM2-7 ring changes from a flat conformation to a right-handed spiral after DNA binding, maintaining in both states a practically flat N-terminal region (Costa et al. 2014). A recent study showed that in a recombinant human MCM2-7 complex, the binding of DNA induces conformational changes including the protrusion of a mass emerging from the AAA^+ region of the helicase (Hesketh et al. 2015); the structure of this active *Hs*MCM2-7 + DNA complex shows a closed, N-terminal planar-ring configuration, where the prominent C-terminal extension is suspected to be the C-terminus of some of the subunits (Hesketh et al. 2015).

Summarizing, the following are some of the new contributions of this work to MCMs´ helicase research field:

1. The structure of the trimeric MCM oligomers resembles that of half MCM hexameric ring, suggesting that the spatial organization of the monomers within the trimer already have the general architecture observed in the hexameric assembly.

2. There is a conformational state where there is a connecting bridge between two MCM trimers, which localize at the putative N-terminal region of the complex, arising near the middle protomer from each side, both putatively corresponding to MCM4 subunits. We propose that this connection could be responsible for the incipient dimerization of the trimers. Nevertheless, we cannot discard the possibility of additional functional roles that could be related with this conformation.

3. Putative HTH MCM6 domains show drastic positional changes, going from recessed localizations to different extended positions, some of which project toward the central channel (*Class 3* map). We propose that HTH motif from MCM6 could work as a structural regulator for DNA for entry through the helicase pore by making a steric impediment, for example, when the conformational state of the ring is not yet a “mature” active one.

4. The sealing of the N-terminal pore of the central channel in the more compacted ring closed conformation (*Class 5* map) could be a specific mechanism for avoiding linear DNA to access the

inside of the helicase. This could be indicative that the conformation of this ring is not yet a "mature" active one.

5. We report a putative interaction between the HTH motifs from the two copies of MCM6 within the helicase (maps from *Class 1*, *Class 4*, *Class 5* and *Class 6*). We propose that this interaction plays an important role on the establishment of a potentially active architecture of MCM4/6/7x2. Additionally, it is plausible that the big mass resulting from this HTH-HTH interaction, which localized at an interface of the ring that seems to be more prone to get open in the *Class 1* map, could work as a bulky clasp. We speculate that this clasp could be transiently displaced, promoting the formation of a gapped ring for allowing the helicase to load into DNA, and would be closed again to capture the DNA molecule during helicase translocation.

6. In a proposed possibly active ("mature") conformation of the MCM4/6/7x2 ring (*Class 1* map), the C-terminal domain of one of the two putative MCM4 protomers forms a finger-like structure extended toward the central channel, reaching a position that would allow it to potentially interact with any DNA molecule inside the ring, suggesting a possible important role for this structure.

7. We propose that the mostly planar state observed in all the different ring conformations, and the ability to get closed by itself, could be an explanation for the intrinsic helicase activity of MCM4/6/7x2. This contrast with the tendency of MCM2-7 to form inactive opened non-planar rings, and its requirement for additional factors to acquire a planar configuration and sealing its gap, which seems to promote the activation of its helicase activity.

A further computational analysis using normal mode analysis (NMA) of the six different conformational states (*Class 1* to *Class 6* 3D maps) of the ring-like MCM4/6/7x2 complex (Figure 67), showed two main groups containing the more structurally similar maps. In one group there were included maps *Class 2, 3* and *4*, and in the second group there were maps *Class 1, 5* and *6*.

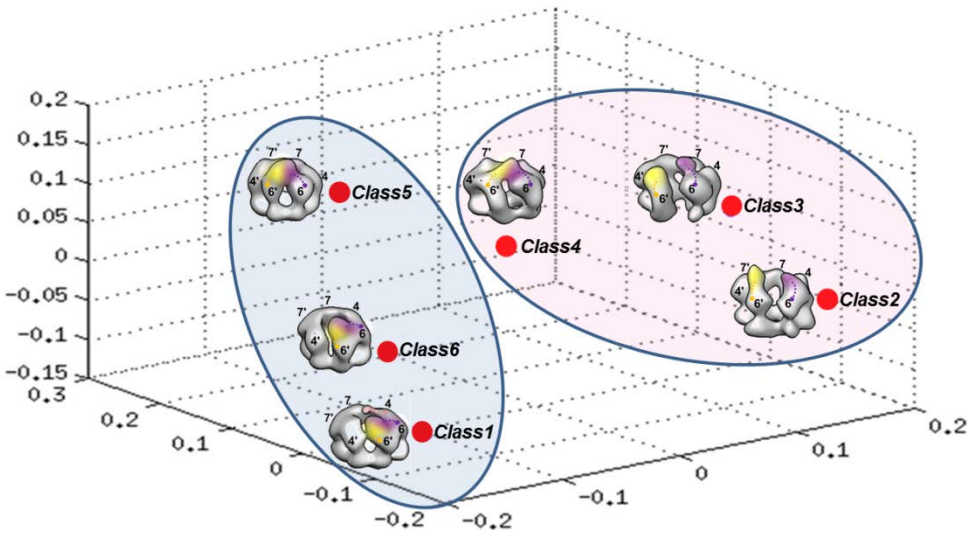


Figure 67. Normal mode analysis (NMA) of the MCM4/6/7x2 ring-like structures *Class 1* to *Class 6*. Graphical representation showing the relative distance correlation (Red dots) among all the analyzed maps, as projected in a 3D space. Blue and red shadowed figures highlight two groups, each containing the more closely related structures. Numbers at the three different axes of the graphic, represent relative values, instead of absolute.

In the following we are going to propose a tentative time sequence of the different EM maps reported in this work (Figure 68):

1. MCM4/6/7 timers are formed. We speculate that two different configurations, MCM7→4→6 and MCM6'→4'→7', could coexist.
2. The two putative trimers, MCM7→4→6 and MCM6'→4'→7', dimerize by making an initial contact through a bridge between the C-terminal domains of the MCM4-4' subunits.
3. The connected trimers approximate by one edge, so that MCM7-7' protomers bind side by side. In this way an open ring with a gap is created at the MCM6-6' interface. The HTH motif of MCM6 is extended toward the central channel making a partial obstruction of the pore, which could work as a steric impediment for DNA passing through at this stage.
4. The N-terminals of both MCM6-6' make contact, closing the lower part of the ring, while their HTH motifs protrude, in different degrees, from the top of the ring.
5. The HTH parts of the MCM6-6' protomers make contact, creating a loose junction that partially obstructs the opening at the top of the ring.
6. Two HTH motifs coming from the MCM6-6' subunits, merge together creating a lump at the top of the ring, in the middle point between the respective two subunits. It would be expected that the interaction of the flexible C-terminal domains stabilizes their HTH fold in a more compact architecture, strengthening the whole union between the MCM6-6'

protomers. At this stage the ring is completely closed and the pore at the N-terminal region is practically sealed, preventing any potential DNA molecule to pass through.

7. The AAA^+ domains from MCM4' and MCM6' separate, giving rise to a notched ring, where the HTH MCM6-6' lump had moved to the top of MCM6, near to the nick.
8. Now that the new interactions between the MCM6-6' subunits had been correctly established within the ring (MCM4'→7'→7→4→6→6'), the lump of the two interacting HTH motifs shifts its position by moving over the notch, making slight contact with the AAA^+ domain of MCM4' such that the nick gets closed again. The central channel is open by both sides but the big aperture at the AAA^+ region is partially divided into two smaller chambers by the putative MCM4 C-terminal finger-like structure projected towards the central pore. In this position the extended C-terminal structure could potentially get in close contact with any DNA crossing through the helicase channel. The pores placed parallel to the central channel, are all big enough for allowing DNA to pass through. The new structural rearrangement of intersubunit contacts in this ring could potentially give rise to a “mature” conformation able to load itself onto DNA and support the helicase activity.

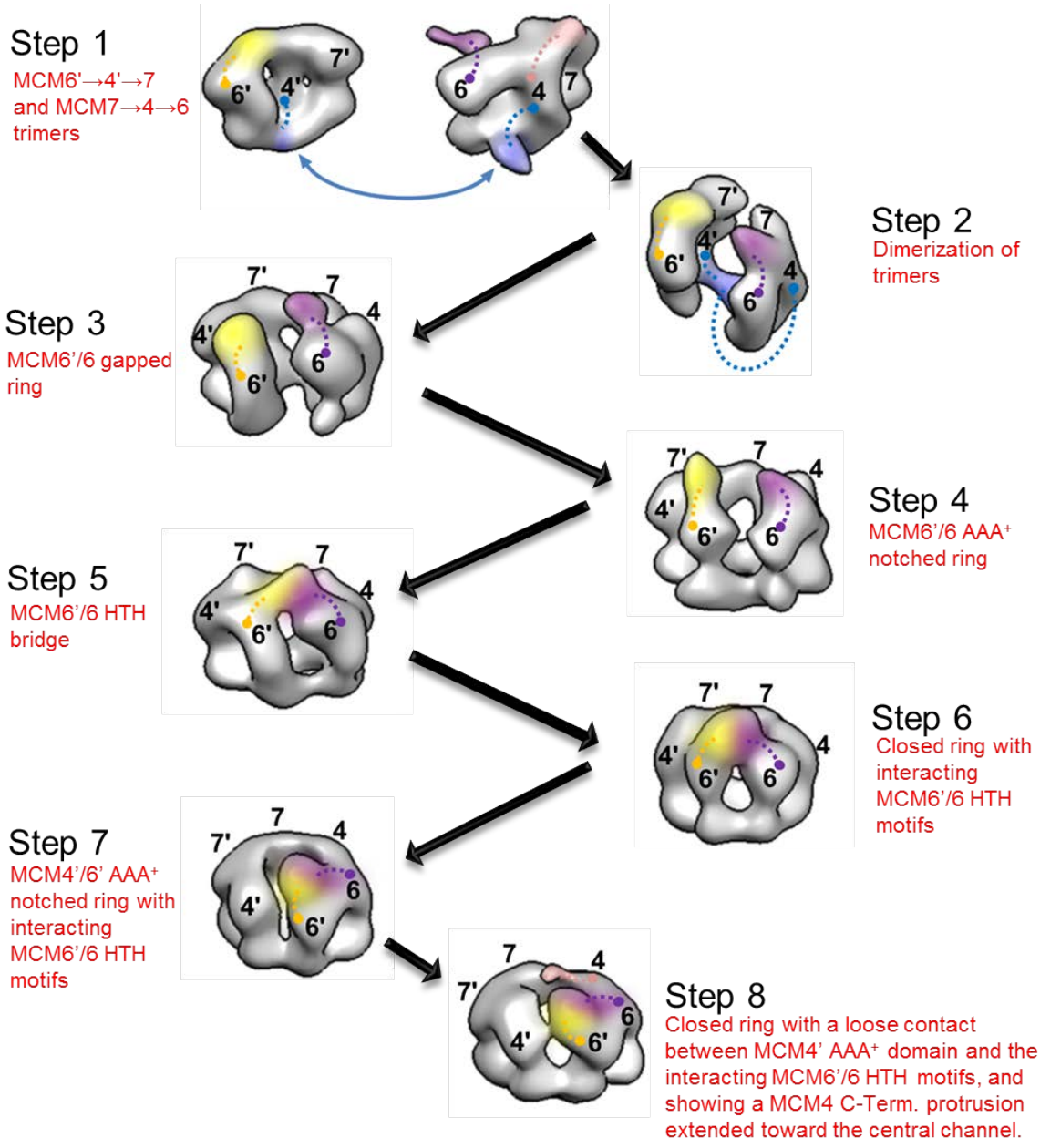


Figure 68. Proposed sequence of a putative multiple-step assembly process of MCM4/6/7x2. The sequence begins with the dimerization of the trimers followed by a number of intermediate conformational changes of the hexamer in a way that protruding HTH domains of both MCM6 and MCM6' interact very closely, and the C-terminal domain of MCM4 protrudes toward the central channel of the helicase. Numbers represent the putative organization of the subunits. Some putatively designated structural features had been highlighted as follows: MCM6' HTH domain (Yellow); MCM6 HTH domain (Purple); N-terminal extension of both MCM4 and MCM4' (either separated or interacting) (Blue), and C-terminal domain of MCM4 (pale-pink). Colored doted-lines represent a symbolic linkage between the respective colored domain and the putative subunit to which it belongs.

Our tentatively proposed sequence of conformational changes provides a structural-based model for the formation of a potentially active MCM4/6/7x2 subunit arrangement (MCM4'→7'→7→4→6→6'), which could somehow emulates the general organization of MCM7→4→6 within the helicase MCM2-7 (MCM5→3→7→4→6→2). We have shown that in the presence of ATP, which might be potentially hydrolyzed, coexist different conformations of hexamers, trimers and monomers of the proteins MCM4, 6 and 7. Which would be the nucleotide binding state (ATP, ADP or *apo*) of each of the presented conformations is a question that still must be answered. Future work on EM analysis of samples in presence of non-hydrolyzable ATP and ADP analogs would help to solve the question. Meanwhile, possible clues can be found on previous studies of MCM. The observation that under *apo* condition mostly trimeric and smaller conformation of MCM appear, while the presence of nucleotide promotes the hexamerizaion, (Ma et al. 2010), strongly points to trimers as being *apo* states while hexamers represent different nucleotide-binding states. When compared with the *apo* state, ATPγS bound *Ecu*MCM2-7 and *Dme*MCM2-7 open rings presents a more compact shape (Lyubimov et al. 2012). These observation would suggest that the closed and, perhaps, even the more compact configurations of MCM4/6/7x2, could correspond to ATP bound states, while the more distended structures could be a mixture of ATP/ADP-binding states.

Studies on *Dme*MCM2-7 (Costa et al. 2011) and *Ecu*MCM2-7 (Lyubimov et al. 2012) have shown that under both conditions, absence or presence of nucleotide, the most prominent conformational state is a non-planar open-ring configuration. It has been proposed that MCM2-7 needs additional factors for closing (through the use of specific anions, as Glutamate or Acetate) (Bochman & Schwacha 2008), (Hesketh et al. 2015) or sealing (making a complex with GINS and Cdc45) (Costa et al. 2011) its MCM2/5 gate, in order to function as an helicase (O'Shea, V.L. and Berger 2014). Interestingly, planar states of MCM2-7 had only been reported in a minority population of partially closed ring particles (*apo* notched *Dme*MCM2-7 (Costa et al. 2011)); in an helicase-DNA complex in presence of specific ions (*Hs*MCM2-7 with helicase activity (Hesketh et al. 2015), or when the helicase is in complex with proteins implicated in its DNA loading (*Sc*OCCM (Sun et al. 2013), (Sun et al. 2014)) or dsDNA unwinding activation (*Dme*CMG (Costa et al. 2011) and). This led us to tentatively propose that a planar state could be required for a correct structural/functional alignment of the subunits, in order to carry out the activation of the helicase activity. Our study demonstrates that mouse helicase MCM4/6/7x2 forms, by itself, mostly N-

terminal planar configurations of the open, notched or completely closed ring, which would give a structural support to its already known intrinsic DNA-unwinding activity (You et al. 1999).

The extensive variety of MCMs complexes that have been identified throughout this work and several others put in evidence a dynamic and highly flexible group of proteins that can be associated with the functional complexity of the eukaryotic replication and transcription regulatory processes, in which they had been found to play important roles. The ensemble of maps presented in this study is, perhaps, just a small but representative sample of what should be a whole range of continuous *apo* and ATP-binding/hydrolyzed conformational states. These structures represent an important progress toward the understanding of the mechanistic behavior of MCM4/6/7x2 helicase and, to some extent, of MCM2-7 and other helicase complexes.

5.4. Structural Characterization of TWINKLE, the human mitochondrial DNA helicase: flexibility and heterogeneity

The ring-shaped mitochondrial replicative DNA helicase Twinkle is an essential component of the minimal replisome required for the correct maintenance of mtDNA. Mutations in the Twinkle gene and mtDNA deletions are both associated with adPEO syndrome. In addition to its DNA unwinding activity, Twinkle presents an antagonistic annealing activity, so it is thought that this enzyme could be implicated in different aspects of DNA processing aside from replication.

Both hexameric and heptameric complexes of Twinkle have been previously reported. Here, we provided further conformational insights using EM techniques that allowed us to identify, aside from the more commonly found hexamers and heptamers, novel octameric and pentameric states of the protein. We described a wide number of conformations of Twinkle where protomers behave as flexible arms, showing different axial displacements and bending movements of their NTD. Flexibility of NTD give rise to multiple forms of both extended and more compacted complexes, agreeing with previous dynamic studies of Twinkle protein in solution. Additionally, our structural comparative analysis between the full-length Twinkle and a truncated construct lacking the ZBD of the protein puts in evidence a notable role of ZBD on the ability of the enzyme to acquire a wide number of conformational states that in absence of this domain do not occur. Our structural characterization of Twinkle showing its highly dynamic and flexible performance can be associated with the structural/functional complexity required for both DNA unwinding and annealing activities.

5.4.1. Effect on the helicase sample to changes in ionic environment

Function and structure of mitochondria are altered by changes in osmotic and ionic conditions of both its internal and external environment, so variations of these factors within the mitochondria

matrix, being importantly influenced by the metabolic state, are likely related to regulation of mitochondria functions (Hackenbrock 1968), (Bradshaw & Pfeiffer 2006), (Murphy & A. Eisner 2010). The precise mechanisms that allow synchronization of cell proliferation with mitochondrial genome replication are still far from being completely understood, but ionic conditions seem to play an essential role. To date, important effects on the state and function of proteins in solution caused by changes in factors such as ionic strength, temperature and molecular crowding (Schroer et al. 2012), (Kuznetsova et al. 2014) have been widely described. Specifically, studies of the dynamic effect of salt concentration on the state of Twinkle protein showed its strong susceptibility to aggregate at concentrations below 250 mM NaCl, especially in the absence of additional stabilizing cofactors (Ziebarth et al. 2010). In agreement with the aforementioned study, we observed that Twinkle is completely soluble at concentrations of 250 mM NaCl or higher, while an important aggregation occurred at 150 mM NaCl, which was much more pronounced at 100 mM NaCl, even in presence of those cofactors that had been described to increase the solubility at these lower salt conditions (Ziebarth et al. 2010). Complete recovery of soluble helicase particles by increasing salt concentration to 330 mM NaCl from partially insoluble samples at 100 mM NaCl plus cofactors, suggests that at the analyzed low salt conditions the oligomers are not dissociated, instead just a tight conglomeration of still assembled complexes could be occurring in response to the buffer conditions. If the latter is true, perhaps this could be used as a mechanism of mitochondria to regulate the availability of Twinkle by using a "sequestration" strategy regulated by changes in the ionic strength of the environment, which in turn would affect mtDNA replication.

5.4.2. Multiple oligomer states of the enzyme

Twinkle is able to form a number of oligomeric complexes which are predominately hexameric and heptameric forms, although we also detected a new population of octamers and pentamers. The diameter of central channel in all the different complexes can easily accommodate ssDNA and, with exception of the pentameric form, the rest could also accommodate a dsDNA molecule. We observed that the population of hexamers was around half of the particles in both full-length and ΔZBD samples, while the proportion of other oligomeric states was smaller and vary significantly between the two proteins. The above suggests that the hexameric configuration could be the most

stable form of Twinkle, which would not be surprising considering that many other different ring-like helicases tend to be organized as hexamers (Patel & Picha 2000).

5.4.3. Structural flexibility, possible opening of the ring and effect of the ZBD

Our NS-EM image analysis revealed that hexameric and heptameric complexes formed by the full-length Twinkle present a wide range of mostly asymmetric conformations produced by multiple combinations of axial displacement and bending of the protomers. The full-length protein sample showed an important number of the n-1 bent arms form, but no particles with all arms contracted were detected. Interestingly, the ΔZBD protein presented both n-1 bent arms extended and fully contracted particles, demonstrating that ZBD is not required to acquire these conformations and suggesting that, instead, RPD/CTD and/or RPD-CTD linker/CTD interactions could likely be involved. On the other hand, the absence of completely contracted conformations in the full-length Twinkle suggests that the presence of ZBD somehow precludes the simultaneous bending of all the protomers, probably by a steric impediment. Considering the absence of fully contracted particle states in the wild type protein and, instead, the presence of a significant proportion of n-1 arms bent particles, it seems important to consider whether there is any functional advantage for the enzyme to acquire a conformation where only one NTD is extended, fully exposed to the solvent. It has been proposed a ring-opening mechanism for ssDNA binding in the central channel of the T7 gp4 (Ahnert et al. 2000) where a transient binding of NTD to a DNA molecule produces conformational changes that open the ring. One possibility would be that in Twinkle, multiple simultaneously NTD-DNA interactions could interfere with the opening of the ring, so only one NTD should initially bind to DNA to correctly activate the helicase opening for its subsequent loading onto the DNA. If only one DNA-binding site is available (the one of the extended monomer) that would restrict the binding of multiple DNA molecules to the helicase. Additionally, a full exposition of only one NTD would help to avoid an initial competition for binding the same DNA molecule by multiple protomers. In the same line of the putative ability of Twinkle to acquire an open conformation, our EM data of the wild type protein showed a small group of particles with a thin gap, which is also suggested in all our 3D initial map reconstructions and in a previous structural report of Twinkle (Fernández-Millán et al. 2015). The low frequency of open-ring

particles observed could be reflecting the stability of closed conformations in absence of DNA. It would be interesting to know whether the presence of DNA promote an increase in the number of particles in open conformation; unfortunately, our attempts to obtain a soluble sample at the low salt concentration that the enzyme require to bind DNA, have failed.

The particles of the wild type protein showed a tendency to arrange some of their extended protomers into closer pairs, which could be in agreement with the report of highly stable dimeric forms of Twinkle resistant to the denaturant conditions of SDS-PAGE (Ziebarth et al. 2010). This observation suggests that multimerization of stable, preformed dimers, could contribute to the formation of more stable ring-like higher order species of Twinkle. In support of this idea, it was proposed in T7 gp4 and T4 gp41, an oligomerization mechanism (Notarnicola et al. 1995), (Dong et al. 1995) where, first, protein dimers are formed through N-terminal "head-to-head" interactions and, subsequently, the new formed dimers oligomerize through C-terminal "tail-to-tail" interactions to form hexameric complexes. All the above-mentioned points would imply the presence of different protein-protein interfaces within the helicase ring. In contrast to the broad number of mostly asymmetric particles formed by the full-length Twinkle protein, the ΔZBD construct showed, aside from the more compacted states, particles with a predominantly symmetric distribution of extended protomers, where the NTDs are isolated from each other.

Contacts of Twinkle *in trans* between ZBDs and RPDs, and between RPDs and CTDs of adjacent protomers are suggested by the fitting of a Twinkle homology model (based on both T7 gp4 and DnaG) into the EM map of a chemically fixed, compact hexameric state of the enzyme (Fernández-Millán et al. 2015). Some of the ZBDs within the EM map were not detected, presumably due to their higher flexibility, which could explain why the putative ZBDs in our EM images of Twinkle are so difficult to detect (Figure 56). Additionally, both *cis* and *trans* interactions between ZBD and RPD are reported in T7 gp4 (Lee & Richardson 2002) and DnaG (Corn et al. 2005); a putative tight interaction *in cis* between ZBD and RPD of Twinkle (Figure 69 B) could also make difficult to identify the ZBD by NS-EM. The aforementioned domain interactions are all compatible with our EM results, which allowed us to propose their possible role on the different conformations observed, as follows: both fully compact (Figure 59 C) and n-1 bent arms (Figure 55 B and Figure 59 B) particle conformations could be produced by *trans* RPD/CTD (Figure 69 C) interactions, while *trans* ZBD/RPD interactions between either two neighbouring (Figure 69 D) or nonconsecutive (Figure 69 E) subunits seem to be responsible for the asymmetric conformations exhibited by the

full-length Twinkle (Figure 55 C), which are not displayed by the ΔZBD protein. Our EM analysis shows that Twinkle´s ZBD is required for such additional conformations observed in the full-length protein, putting in evidence an essential role of this domain for the naturally highly dynamic behavior of the enzyme.

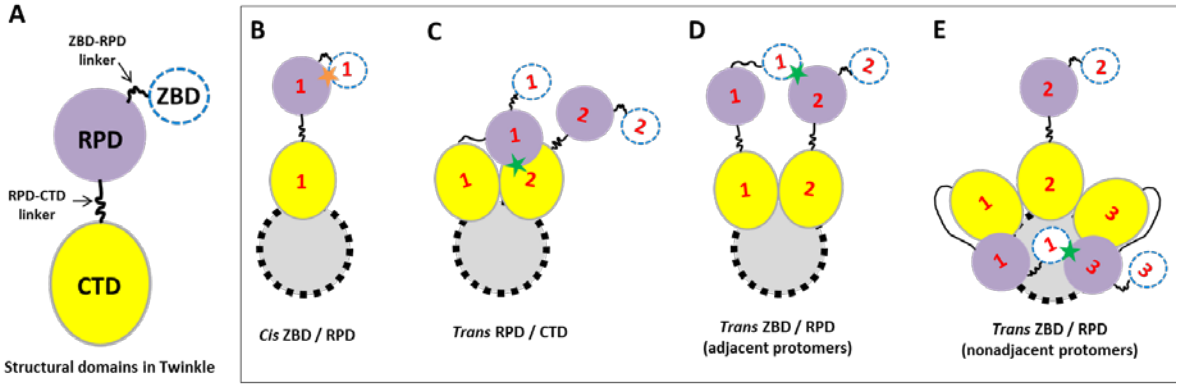


Figure 69. Putative domain interactions of Twinkle that could explain the conformations observed by NS-EM. (A) Representation of structural domains of Twinkle. CTD is shown in yellow and RPD in purple. ZBD is shown as a blue dashed circle, meaning that is not always detected by EM. The RPD-CTD and ZBD-RPD linkers are shown as wavy black lines. **(B)** *Cis* ZBD/RPD interaction. **(C)** *Trans* RPD/CTD and **(D)** ZBD/CTD interactions between two consecutive protomers. **(E)** *Trans* ZBD/RPD interactions between two nonconsecutive subunits. *Cis* and *trans* contacts between domain are represented by an orange and green star, respectively. In figures **B** to **E,** the gray circle delineated by a black dashed line represents the central channel of a ring-shaped complex; for simplifying the scheme, there is shown the minimal number of subunits that is required to explain each of the different types of domain interactions. Domains belonging to one subunit are labeled with the same number.

Taking together, with the current knowledge about Twinkle, which include the identification of several highly flexible oligomeric states, the different intra and intersubunit interactions, the ability to carry out the antagonist helicase and unwinding activities (in a NTPase-dependent or NTP-independent manner, respectively), the presence of three different DNA binding sites and its affinity for binding different single-stranded and double-stranded DNA substrates, it seems likely that Twinkle plays a role in different aspects of DNA processing beyond replication. It has been reported that although truncation of the ZBD decreases the ability of the enzyme to bind ssDNA which, in consequence, also affects its ssDNA-stimulation of the ATPase activity, it does not significantly alters the activity of the replisome *in vitro* (Farge et al. 2008). Our results showing a notable structural role of the ZBD on the ability of the enzyme to acquire a wide number or conformational changes that, in absence of this domain do not occur, suggest that there must be an important, although not yet described functional role for these interdomain interactions in Twinkle´s activity. Since deletion of the ZBD does not seems to importantly affect the helicase activity of Twinkle, it could be playing a more essential role in other activities of the enzyme such

as DNA annealing ability (Sen et al. 2012), (Wu 2012), or in other regulatory activities under *in vivo* conditions. Our observations raise important new questions that, we expect, will lead the foundation for future research exploring further structural features and functional roles of Twinkle´s flexibility and its ZBD.

Chapter 6

6. Conclusions

Concerning human CPAP subproject:

1. CPAP is a highly versatile protein essential for centriole biogenesis, reported to interact with many different proteins along the cell cycle. Together, the biochemical, biophysical and structural studies carried out allowed, for the first time, the isolation and identification of multiple homomeric complexes formed by $CPAP^{897-1338}$. Structural characterization of the different forms of the protein showed that, in its monomeric state, $CPAP^{897-1338}$ behaves as a flexible filament, while the three dimensional reconstructions of both a dimeric and a tetrameric assemblies of the protein, revealed globular and well-structured conformations.

2. The analysis of purified modular column-like structures with an axial repetition pattern around 8 nm, suggested that they are formed by stacks of the $CPAP^{897-1338}$ tetramer, and, furthermore, that they could correspond to similar structures observed at the inner walls of the centriole.

Concerning mouse MCM4/6/7 subproject:

3. MCM4/6/7 helicase has a dynamic structural behavior in presence of ATP. The three dimensional analysis of the heteromeric MCM4/6/7 sample revealed, for the first time, the structure of different complexes that include: trimers, what seems to be the incipient dimerization of trimers and several ring-like hexameric conformations.

4. Fitting of homologue atomic structures of MCM proteins into the obtained three dimensional maps, allowed the putative localization of structural domains, showing especially flexible regions, putatively designated to either separate or interacting C-

terminus of the two MCM6 copies in the helicase ring assemblies. Additionally, a small finger-like structure in one of the maps could correspond to the C-terminal domain of one of the two MCM4 protomers in the complex.

5. Supported by previous biochemical and structural studies of MCM helicases, structural features of our MCM4/6/7 density maps suggest that the subunit organization within the helicase hexamer is: MCM4'→MCM7'→MCM7→MCM4→MCM6→MCM6'

Concerning human Twinkle subproject:

6. The homomeric helicase Twinkle forms highly flexible assemblies, primarily hexamers and heptamers, although pentamers and octamers are found in a minority proportion.

7. Differently extended or compact conformations of Twinkle result as consequence of multiple combinations of the axial movements and bending of the monomers.

8. Comparison between the full-length protein and a N-terminal deletion mutant showed a significative structural role of the zinc binding domain on the ability of the enzyme to acquire a wide number or conformational changes.

9. Initial three dimensional reconstructions showed that extended conformations of the helicase are relatively flat, while the compact ones present a more bulky structure suggesting the presence of a second floor.

Conclusiones

En referencia al subproyecto de CPAP:

1. CPAP en una proteína versátil, esencial en la biogénesis centriolar, la cuál se ha reportado que interacciona con distintas proteínas a lo largo del ciclo celular. En su conjunto, los estudios bioquímicos, biofísicos y estructurales que se realizaron, han permitido, por primera vez, la identificación y el aislamiento de múltiples complejos de oligómeros formados por $CPAP^{897-1338}$. La caracterización estructural de las distintas formas en que se encuentra la proteína, mostraron que en su estado monomérico $CPAP^{897-1338}$ se comporta como hebras flexibles, mientras que las reconstrucciones en tres dimensiones de la forma dimérica y tetramérica, revelaron conformaciones globulares bien estructurales.
2. El análisis de unas estructuras modulares en forma de columna y con un patrón de repetición axial cercano a los 8 nm, sugieren que dichas estructuras están formadas por pilas de tetrámeros de $CPAP^{897-1338}$, y además, que podrían corresponder a estructuras similares que se encuentran localizadas en las paredes internas del centriolo.

En referencia al subproyecto de MCM4/6/7:

3. La helicasa MCM/6/7 muestra un comportamiento estructural muy dinámico en presencia de ATP. El análisis en tres dimensiones de la muestra heteromérica MCM4/6/7 permitió mostrar, por primera vez, la estructura de diferentes complejos, entre los cuales se incluyen: trímeros, lo que pareciese ser un estadio incipiente de dimerización de los trímeros, y varias conformaciones de hexámeros en forma de anillo.
4. El ajuste de modelos atómicos por homología dentro de los mapas en tres dimensiones obtenidos, permitió localizar dominios estructurales de la proteína de manera putativa, mostrando zonas especialmente flexibles que fueron asignadas, también en forma putativa, a la región C-terminal de las dos copias de MCM6 que forman parte del complejo helicasa. Adicionalmente, se identificó una estructura en forma de dedo que podría corresponder a la región del C-terminal de una de las dos copias de MCM4.

5. Apoyándose en previos estudios bioquímicos y estructurales de MCM, varios de los elementos estructurales en nuestros mapas de densidad sugieren la siguiente organización de los protómeros dentro del complejo MCM4/6/7: MCM4'→MCM7'→MCM7→MCM4→MCM6→MCM6'

En referencia al subproyecto de Twinkle:

6. Twinkle forma complejos altamente flexibles que, es su mayoría, corresponden a formas hexaméricas y heptaméricas, aunque también se encuentra un número minoritario de pentámeros y octámeros.
7. El movimiento de los protómeros de la helicasa en distintas direcciones, da lugar a múltiples conformaciones extendidas o compactas.
8. La comparación entre la proteína completa y un mutante con una dilección en la región N-terminal de Twinkle, muestra que el dominio de unión a zinc juega un papel estructural importante para que la enzima adquiera un amplio número de cambios conformacionales.
9. Las diferentes reconstrucciones iniciales en tres dimensiones que se obtuvieron, muestran que las conformaciones extendidas de Twinkle son relativamente planas, mientras que las compactas presentan una estructura más abultada que sugiere la presencia de un segundo piso.

References

Adachi, Y., Usukura, J. & Yanagida, M., 1997. A globular complex formation by Nda1 and the other five members of the MCM protein family in fission yeast. *Genes to Cells*, 2(7), pp.467–479.

Ahnert, P., Picha, K.M. & Patel, S.S., 2000. A ring-opening mechanism for DNA binding in the central channel of the T7 helicase - primase protein. *EMBO*, 19(13), pp.3418–27.

Alves-Cruzeiro, J.M.D.C., Nogales-Cadenas, R. & Pascual-Montano, A.D., 2014. CentrosomeDB: a new generation of the centrosomal proteins database for Human and Drosophila melanogaster. *Nucleic acids research*, 42(Database issue), pp.D430–6.

Amos, L.A., Henderson, R. & Unwln, P.N.T., 1982. Three-dimensional structure determination by electron microscopy of two-dimensional crystals. *Progress in Biophysics and Molecular Biology*, 39, pp.183–231.

Aravind, L. & Koonin, E. V, 1999. DNA-binding proteins and evolution of transcription regulation in the archaea. *Nucleic acids research*, 27(23), pp.4658–4670.

Aravind, L., Leipe, D.D. & Koonin, E. V, 1998. Toprim - a conserved catalytic domain in type IA and II topoisomerases , DnaG-type primases , OLD family nucleases and RecR proteins. *Nucleic Acids Research*, 26(18), pp.4205–4213.

Arias-Palomo, E. et al., 2013. The bacterial DnaC helicase loader is a DnaB ring breaker. *Cell*, 153(2), pp.438–48.

Azimzadeh, J. & Bornens, M., 2007. Structure and duplication of the centrosome. *Journal of cell science*, 120(Pt 13), pp.2139–42.

Bacteriophage, T. et al., 1995. Nucleotide and DNA-induced Conformational Changes in the. *Journal of Biological Chemistry*, 270(41), pp.24509–24517.

Bae, B. et al., 2009. Insights into the Architecture of the Replicative Helicase from the Structure of an Archaeal MCM Homolog. *Structure/Folding and Design*, 17(2), pp.211–222.

Barry, E.R. et al., 2007. Archaeal MCM has separable processivity, substrate choice and helicase domains. *Nucleic acids research*, 35(3), pp.988–98.

Barry, E.R. et al., 2009. Intersubunit allosteric communication mediated by a conserved loop in the MCM helicase. *Proceedings of the National Academy of Sciences of the United States of America*, 106(4), pp.1051–6.

Bell, S.D. & Botchan, M.R., 2013. The minichromosome maintenance replicative helicase. *Cold Spring Harbor perspectives in biology*, 5(11), p.a012807.

Bettencourt-Dias, M. & Glover, D.M., 2007. Centrosome biogenesis and function: centrosomics brings new understanding. *Nature reviews. Molecular cell biology*, 8(6), pp.451–63.

Biswas-Fiss, E., Khopde, S. & Biswas, S., 2005. The Mcm467 complex of Saccharomyces cerevisiae is preferentially activated by autonomously replicating DNA sequences. *Biochemistry*, pp.2916–2925.

Bochman, M.L., Bell, S.P. & Schwacha, A., 2008. Subunit organization of Mcm2-7 and the unequal role of active sites in ATP hydrolysis and viability. *Molecular and cellular biology*, 28(19), pp.5865–73.

Bochman, M.L. & Schwacha, A., 2007. Differences in the single-stranded DNA binding activities of MCM2-7 and MCM467: MCM2 and MCM5 define a slow ATP-dependent step. *The Journal of biological chemistry*, 282(46), pp.33795–804.

Bochman, M.L. & Schwacha, A., 2009. The Mcm complex: unwinding the mechanism of a replicative helicase. *Microbiology and molecular biology reviews : MMBR*, 73(4), pp.652–83.

Bochman, M.L. & Schwacha, A., 2008. The Mcm2-7 complex has in vitro helicase activity. *Molecular cell*, 31(2), pp.287–93.

Bradshaw, P.C. & Pfeiffer, D.R., 2006. Release of Ca2+ and Mg2+ from yeast mitochondria is stimulated by increased ionic strength. *BMC biochemistry*, 12(February), pp.1–12.

Brewster, A.S. et al., 2008. Crystal structure of a near-full-length archaeal MCM : Functional insights for an AAA hexameric helicase. *Proceedings of the National Academy of Sciences*, 105(51), pp.20191–20196.

Brewster, A.S. & Chen, X.S., 2010. Insights into MCM functional mechanism: lessons learned from the archaeal MCM complex. *Critical reviews in biochemistry and molecular biology*, 45(3), pp.243–256.

Burger, G., Gray, M.W. & Lang, B.F., 2003. Mitochondrial genomes : anything goes. *Trends in cell biology*, 19(12), pp.709–716.

Burgess, S. a et al., 2004. Use of negative stain and single-particle image processing to explore dynamic properties of flexible macromolecules. *Journal of structural biology*, 147(3), pp.247–58.

Burkhart, R., Schulte, D. & Hu, B., 1995. Interactions of human nuclear proteins P1Mcm3 and P1Cdc46. *European Journal of Biochemistry*, 438(2), pp.431–438.

Carvalho-Santos, Z. et al., 2010. Stepwise evolution of the centriole-assembly pathway. *Journal of cell science*, 123(Pt 9), pp.1414–26.

Chang, J. et al., 2010. PLK2 phosphorylation is critical for CPAP function in procentriole formation during the centrosome cycle. *The EMBO journal*, 29(14), pp.2395–406.

Chen, C.-Y. et al., 2006. CPAP interacts with 14-3-3 in a cell cycle-dependent manner. *Biochemical and Biophysical Research Communications*, 342(4), pp.1203–1210.

Chen, Y., Wei, L. & Mu, J.D., 2003. Probing protein oligomerization in living cells with fluorescence fluctuation spectroscopy. *Proceedings of the National Academy of Sciences*, 100(26), pp.15492–15497.

Chenna, R., 2003. Multiple sequence alignment with the Clustal series of programs. *Nucleic Acids Research*, 31(13), pp.3497–3500.

Cho, J.-H. et al., 2006. Depletion of CPAP by RNAi disrupts centrosome integrity and induces multipolar spindles. *Biochemical and Biophysical Research Communications*, 339(3), pp.742–747.

Chong, J.P.J., Thmmes, P. & Blow, J.J., 1996. The role of MCM / PI proteins in the licensing of DNA replication. *Trends in Biochemical Sciences*, 21(3), pp.102–106.

Cizmecioglu, O. et al., 2010. Cep152 acts as a scaffold for recruitment of Plk4 and CPAP to the centrosome. *The Journal of cell biology*, 191(4), pp.731–9.

Claycomb, J.M. et al., 2002. Visualization of replication initiation and elongation in Drosophila. *The Journal of cell biology*, 159(2), pp.225–36.

Cole, C., Barber, J.D. & Barton, G.J., 2008. The Jpred 3 secondary structure prediction server. *Nucleic acids research*, 36(Web Server issue), pp.W197–201.

Comartin, D. et al., 2013. CEP120 and SPICE1 cooperate with CPAP in centriole elongation. *Current biology : CB*, 23(14), pp.1360–6.

Cong, Y. & Ludtke, S.J., 2010. *Single Particle Analysis at High Resolution* 1st ed., Elsevier Inc.

Cormier, A. et al., 2009. The PN2-3 domain of centrosomal P4.1-associated protein implements a novel mechanism for tubulin sequestration. *The Journal of biological chemistry*, 284(11), pp.6909–17.

Corn, J.E. et al., 2005. Crosstalk between Primase Subunits Can Act to Regulate Primer Synthesis in trans. *Molecular Cell*, 20(3), pp.391–401.

Costa, A. et al., 2014. DNA binding polarity, dimerization, and ATPase ring remodeling in the CMG helicase of the eukaryotic replisome. *eLife*, 3, p.e03273.

Costa, A. et al., 2006a. Structural basis of the Methanothermobacter thermautotrophicus MCM helicase activity. *Nucleic acids research*, 34(20), pp.5829–38.

Costa, A. et al., 2006b. Structural studies of the archaeal MCM complex in different functional states. *Journal of structural biology*, 156(1), pp.210–9.

Costa, A. et al., 2011. The structural basis for MCM2-7 helicase activation by GINS and Cdc45. *Nature structural & molecular biology*, 18(4), pp.471–7.

Costa, A. & Onesti, S., 2009. Structural biology of MCM helicases. *Critical reviews in biochemistry and molecular biology*, 44(5), pp.326–42.

Cottee, M. a et al., 2013. Crystal structures of the CPAP/STIL complex reveal its role in centriole assembly and human microcephaly. *eLife*, 2, p.e01071.

Coué, M., Amariglio, F. & Maiorano, D., 1998. Evidence for Different MCM Subcomplexes with Differential Binding to Chromatin in Xenopus. *Experimental Cell Research*, 289(2), pp.282–289.

Crampton, D.J. et al., 2006. Oligomeric States of Bacteriophage T7 Gene 4 Primase / Helicase. *Journal of Molecular Biology*, 360(3), pp.667–677.

Crevel, G. et al., 2007. Differential requirements for MCM proteins in DNA replication in Drosophila S2 cells. *PloS one*, 2(9), p.e833.

DaFonseca, C.J., Shu, F. & Zhang, J.J., 2001. Identification of two residues in MCM5 critical for the assembly of MCM complexes and Stat1-mediated transcription activation in response to IFN-gamma. *Proceedings of the National Academy of Sciences of the United States of America*, 98(6), pp.3034–9.

Dalton, S. & Whitbread, L., 1995. Cell cycle-regulated nuclear import and export of Cdc47, a protein essential for initiation of DNA replication in budding yeast. *Proceedings of the National*, 92(7), pp.2514–2518.

Daniel, D.C., Dagdanova, A. V & Johnson, E.M., 2013. The MCM and RecQ Helicase Families : Ancient Roles in DNA Replication and Genomic Stability Lead to Distinct Roles in Human Disease. In D. D. Stuart, ed. *The Mechanisms of DNA Replication*. pp. 59–89.

Das, M. et al., 2014. MCM Paradox: Abundance of Eukaryotic Replicative Helicases and Genomic Integrity. *Molecular Biology International*, pp.1–11.

Davey, M.J., Indiani, C. & O'Donnell, M., 2003. Reconstitution of the Mcm2-7p heterohexamer, subunit arrangement, and ATP site architecture. *The Journal of biological chemistry*, 278(7), pp.4491–9.

Debec, A. & Sullivan, W., 2010. Centrioles : active players or passengers during mitosis ? *Cell Molecular Life Science*, 67, pp.2173–2194.

Dimitrova, D. & Todorov, I., 1999. Mcm2, but not RPA, is a component of the mammalian early G1-phase prereplication complex. *The Journal of cell biology*, 146(4), pp.709–722.

Dobbins, S.E., Lesk, V.I. & Sternberg, M.J.E., 2008. Insights into protein flexibility : The relationship between normal modes and conformational change upon protein – protein docking. *Proceedings of the National Academy of Sciences*, 105(30), pp.10390–5.

Dong, F., Gogol, E.P. & Von Hippel, P.H., 1995. The phage T4-coded DNA replication helicase (gp41) forms a hexamer upon activation by nucleoside triphosphate. *Journal of Biological Chemistry*, 270(13), pp.7462–7473.

Doxsey, S., 2001. RE-EVALUATING CENTROSOME FUNCTION. *Nature Reviews Molecular Cell Biology*, 2, pp.688–698.

Doxsey, S., Zimmerman, W. & Mikule, K., 2005. Centrosome control of the cell cycle. *Trends in Cell Biology*, 15(6), pp.303–311.

Dündar, H. et al., 2012. Identification of a novel Twinkle mutation in a family with infantile onset spinocerebellar ataxia by whole exome sequencing. *Pediatric Neurology*, 46(3), p.2012.

Dziak, R. et al., 2003. Evidence for a role of MCM (mini-chromosome maintenance)5 in transcriptional repression of sub-telomeric and Ty-proximal genes in Saccharomyces cerevisiae. *The Journal of biological chemistry*, 278(30), pp.27372–81.

Egli, M., 2010. Diffraction techniques in structural biology. In *Current Protocols in Nucleic Acid Chemistry*.

Ellis, R.J. & Ellis, R.J., 2001. Macromolecular crowding : obvious but underappreciated. *tre*, 26(10), pp.597–604.

Erzberger, J.P. & Berger, J.M., 2006. Evolutionary relationships and structural mechanisms of AAA+ proteins. *Annual review of biophysics and biomolecular structure*, 35, pp.93–114.

Evrin, C. et al., 2009. A double-hexameric MCM2-7 complex is loaded onto origin DNA during licensing of eukaryotic DNA replication. *Proceedings of the National Academy of Sciences of the United States of America*, 106(48), pp.20240–5.

Fan, L. et al., 2006. A Novel Processive Mechanism for DNA Synthesis Revealed by Structure , Modeling and Mutagenesis of the Accessory Subunit of Human Mitochondrial DNA Polymerase. *Journal of molecular biology*, 358(5), pp.1229–1243.

Farge, G. et al., 2008. The N-terminal domain of TWINKLE contributes to single-stranded DNA binding and DNA helicase activities. *Nucleic acids research*, 36(2), pp.393–403.

Fernández-Millán, P. et al., 2015. The hexameric structure of the human mitochondrial replicative helicase Twinkle. *Nucleic acids research*, March, pp.1–12.

Fitch, M.J., Donato, J.J. & Tye, B.K., 2003. Mcm7, a subunit of the presumptive MCM helicase, modulates its own expression in conjunction with Mcm1. *The Journal of biological chemistry*, 278(28), pp.25408–16.

Fletcher, R.J. et al., 2003. The structure and function of MCM from archaeal M. Thermoautotrophicum. *Nature structural biology*, 10(3), pp.160–7.

Forsburg, S.L. et al., 1997. Mutational analysis of Cdc19p, a Schizosaccharomyces pombe MCM protein. *Genetics*, 1041(3), pp.1025–1041.

Frank, J. (Howard H.M.I., 2006. *Three-dimensional Electron Microscopy of Macromolecular Assemblies*, Oxford University Press.

Fratter, C. et al., 2010. The clinical , histochemical , and molecular spectrum of PEO1 (Twinkle)-linked adPEO. *Neurology*, 74(20), pp.1619–1626.

Frueh, D.P. et al., 2013. NMR methods for structural studies of large monomeric and multimeric proteins. *Current opinion in structural biology*, 23(5), pp.734–9.

Giaginis, C. & Vgenopoulou, S., 2010. MCM proteins as diagnostic and prognostic tumor markers in the clinical setting. *Histology and Histopathology*, 25(3), pp.351–370.

Giese, K.C. & Vierling, E., 2002. Changes in oligomerization are essential for the chaperone activity of a small heat shock protein in vivo and in vitro. *The Journal of biological chemistry*, 277(48), pp.46310–8.

Gineau, L., Cognet, C. & Kara, N., 2012. Partial MCM4 deficiency in patients with growth retardation, adrenal insufficiency, and natural killer cell deficiency. *The Journal of Clinical Investigation*, 122(3), pp.821–832.

Glaeser, R.M. & Hall, R.J., 2011. Reaching the information limit in cryo-EM of biological macromolecules: experimental aspects. *Biophysical journal*, 100(10), pp.2331–7.

Gómez, E.B., Catlett, M.G. & Forsburg, S.L., 2002. Different phenotypes in vivo are associated with ATPase motif mutations in Schizosaccharomyces pombe minichromosome maintenance proteins. *Genetics*, 160(April), pp.1305–1318.

Gopalakrishnan, J. et al., 2011. Sas-4 provides a scaffold for cytoplasmic complexes and tethers them in a centrosome. *Nature communications*, 2(May), p.359.

Graham, B.W. et al., 2011. Steric exclusion and wrapping of the excluded DNA strand occurs along discrete external binding paths during MCM helicase unwinding. *Nucleic Acids Research*, 39(15), pp.6585–6595.

Gray, M.W., Burger, G. & Lang, B.F., 1999. Mitochondrial Evolution. *Science*, 283(5407), pp.1476–1482.

Green, M. & Sambrook, J., 2012. *Molecular cloning: a laboratory manual* 4th ed., New York: Cold Spring Harbor Laboratory.

Grigorieff, N., 2007. FREALIGN: high-resolution refinement of single particle structures. *Journal of structural biology*, 157(1), pp.117–25.

Gudi, R. et al., 2014. Centrobin-Centrosomal Protein 4.1-associated Protein (CPAP) Interaction Promotes CPAP Localization to the Centrioles during Centriole Duplication. *The Journal of biological chemistry*, 289(22), pp.15166–15178.

Guo, S., 1999. The Linker Region between the Helicase and Primase Domains of the Bacteriophage T7 Gene 4 Protein Is Critical for Hexamer Formation. *Journal of Biological Chemistry*, 274(42), pp.30303–30309.

Hackenbrock, B.Y.C.R., 1968. Chemical and physical fixation of isolated mitochondria in low-energy and high-energy states. *Proceedings of the National Academy of Sciences*, 61(2), pp.598–605.

Hatzopoulos, G.N. et al., 2013. Structural analysis of the G-box domain of the microcephaly protein CPAP suggests a role in centriole architecture. *Structure (London, England : 1993)*, 21(11), pp.2069–77.

Haven, N., 1994. Domains of Escherichia coli primase : Functional activity of a 47-kDa N-terminal proteolytic fragment. *Proceedings of the National Academy of Sciences*, 91(24), pp.11462–11466.

Havugimana, P.C. et al., 2012. A Census of Human Soluble Protein Complexes. *Cell*, 150(5), pp.1068–1081.

Heel, M. Van & Schatz, M., 2005. Fourier shell correlation threshold criteria q. *Journal of structural biology*, 151(3), pp.250–262.

Henderson, R. et al., 2012. Outcome of the First Electron Microscopy Validation Task Force Meeting. *Structure*, 20(2), pp.205–214.

Hennessy, K.M. et al., 1991. A group of interacting yeast DNA replication genes. *Genes & Development*, 5(6), pp.958–969.

Hesketh, E.L. et al., 2015. DNA induces conformational changes in a recombinant human minichromosome maintenance complex. *Journal of Biological Chemistry*, 290(12), pp.1–15.

Hingorani, M.M. & Patel, S.S., 1993. Interactions of Bacteriophage T7 DNA Primase/Helicase Protein with Single-Stranded and Double-Stranded DNAs. *Biochemistry*, 32(46), pp.12478–12487.

Holthoff, H.P., 1998. Human Protein MCM6 on HeLa Cell Chromatin. *Journal of Biological Chemistry*, 273(13), pp.7320–7325.

Hsu, W.-B. et al., 2008. Functional characterization of the microtubule-binding and -destabilizing domains of CPAP and d-SAS-4. *Experimental Cell Research*, 314(14), pp.2591–2602.

Hu, B. et al., 1993. The P1 family : a new class of nuclear mammalian proteins related to the yeast Mcm replication proteins. *nuc*, 21(23), pp.5289–5293.

Hughes, C.R. et al., 2012. MCM4 mutation causes adrenal failure, short stature, and natural killer cell deficiency in humans. *The Journal of clinical investigation*, 122(3), pp.814–820.

Hung, L. et al., 2004. Identification of a Novel Microtubule-destabilizing Motif in CPAP That Binds to Tubulin Heterodimers and Inhibits Microtubule Assembly □. *Molecular Biology of the Cell*, 15(June), pp.2697–2706.

Hung, L.Y., Tang, C.J. & Tang, T.K., 2000. Protein 4.1 R-135 Interacts with a Novel Centrosomal Protein (CPAP) Which Is Associated with the γ-Tubulin Complex. *Molecular and Cellular Biology*, 20(20), pp.7813–7825.

Hydrolysis, D.A.T.P. et al., 1984. Structural and Functional Studies of the dnaB Protein Using Limited Proteolysis. *Journal of Biological Chemistry*, 259(1), pp.88–96.

Hyrien, O., Marheineke, K. & Goldar, A., 2003. Paradoxes of eukaryotic DNA replication: MCM proteins and the random completion problem. *BioEssays : news and reviews in molecular, cellular and developmental biology*, 25(2), pp.116–25.

Ibarra, A., Schwob, E. & Méndez, J., 2008. Excess MCM proteins protect human cells from replicative stress by licensing backup origins of replication. *Proceedings of the National Academy of Sciences of the United States of America*, 105(26), pp.8956–61.

Ichinose, S., 1996. Binding of Human Minichromosome Maintenance Proteins with Histone H3. *Journal of Biological Chemistry*, 271(39), pp.24115–24122.

Ilves, I. et al., 2010. Activation of the MCM2-7 helicase by association with Cdc45 and GINS proteins. *Molecular cell*, 37(2), pp.247–58.

Ilyina, T. V, Gorbalenya, A.E. & Koonin, E. V, 1992. Organization and Evolution of Bacterial and Bacteriophage Primase-Helicase Systems. *Journal of molecular evolution*, 34(4), pp.351–357.

Ishimi, Y., 1997a. A DNA Helicase Activity Is Associated with an MCM4, -6, and -7 Protein Complex. *Journal of Biological Chemistry*, 272(39), pp.24508–24513.

Ishimi, Y., 1997b. A DNA helicase activity is associated with an MCM4, -6, and -7 protein complex. *The Journal of biological chemistry*, 272(39), pp.24508–13.

Ishimi, Y. et al., 2001. Biochemical activities associated with mouse Mcm2 protein. *The Journal of biological chemistry*, 276(46), pp.42744–52.

Iyer, L.M. et al., 2004. Evolutionary history and higher order classification of AAA+ ATPases. *Journal of structural biology*, 146(1-2), pp.11–31.

Jemt, E. et al., 2011. The mitochondrial DNA helicase TWINKLE can assemble on a closed circular template and support initiation of DNA synthesis. *Nucleic acids research*, 39(21), pp.9238–49.

Jenkinson, E.R. & Chong, J.P.J., 2006. Minichromosome maintenance helicase activity is controlled by N- and C-terminal motifs and requires the ATPase domain helix-2 insert. *Proceedings of the National Academy of Sciences*, 103(20), pp.7613–7618.

Ji, K. et al., 2014. Twinkle mutations in two Chinese families with autosomal dominant progressive external ophthalmoplegia. *Neurological Sciences*, 35(3), pp.443–448.

Jin, Q. et al., 2014. Iterative elastic 3D-to-2D alignment method using normal modes for studying structural dynamics of large macromolecular complexes. *Structure*, 22(3), pp.496–506.

Kanter, D.M., Bruck, I. & Kaplan, D.L., 2008. Mcm subunits can assemble into two different active unwinding complexes. *The Journal of biological chemistry*, 283(45), pp.31172–82.

Kasiviswanathan, R. et al., 2004. Biochemical characterization of the Methanothermobacter thermautotrophicus minichromosome maintenance (MCM) helicase N-terminal domains. *The Journal of biological chemistry*, 279(27), pp.28358–66.

Kato, M. et al., 2003. Modular Architecture of the Bacteriophage T7 Primase Couples RNA Primer Synthesis to DNA Synthesis. *Molecular cell*, 11(5), pp.1349–1360.

Kearsey, S.E. & Labib, K., 1998. MCM proteins: evolution, properties, and role in DNA replication. *Biochimica et biophysica acta*, 1398(2), pp.113–36.

Kearsey, S.E., Labibt, K. & Maiorano, D., 1996. Cell cycle control of eukaryotic DNA replication. *Current opinion in Genetics and Development*, 6(2), pp.208–214.

Keck, J.L. et al., 2000. Structure of the RNA Polymerase Domain of E . coli Primase. *Science*, 287(5462), pp.2482–2487.

Kelman, Z., Lee, J.-K. & Hurwitz, J., 1999. The single minichromosome maintenance protein of Methanobacterium thermoautotrophicum Delta H contains DNA helicase activity. *Proceedings of the National Academy of Sciences*, 96(26), pp.14783–14788.

Kim, M.K., Dudognon, C. & Smith, S., 2012. Tankyrase 1 regulates centrosome function by controlling CPAP stability. *EMBO reports*, 13(8), pp.724–32.

Kimura, H. et al., 1995. Molecular cloning of cDNA encoding mouse Cdc21 and CDC46 homologs and characterization of the products: physical interaction between P1 (MCM3) and CDC46 proteins. *Nucleic Acids Research*, 23(12), pp.2097–2104.

Kinoshita, Y. & Johnson, E.M., 2004. Site-specific loading of an MCM protein complex in a DNA replication initiation zone upstream of the c-MYC gene in the HeLa cell cycle. *The Journal of biological chemistry*, 279(34), pp.35879–89.

Kirkham, M. et al., 2003. SAS-4 is a C. elegans centriolar protein that controls centrosome size. *Cell*, 112(4), pp.575–87.

Kitagawa, D. et al., 2011. Spindle positioning in human cells relies on proper centriole formation and on the microcephaly proteins CPAP and STIL. *Journal of cell science*, 124(Pt 22), pp.3884–93.

Kley, J. et al., 2011. Structural adaptation of the plant protease Deg1 to repair photosystem II during light exposure. *Nature structural & molecular biology*, 18(6), pp.728–31.

Kleylein-sohn, J. et al., 2007. Plk4-induced centriole biogenesis in human cells. *Developmental Cell*, 13(2), pp.190–202.

Kohlmaier, G. et al., 2009. Overly long centrioles and defective cell division upon excess of the SAS-4-related protein CPAP. *Current Biology*, 19(12), pp.1012–1018.

Koonin, E. V, 1993. A common set of conserved motifs in a vast variety of putative nucleic acid-dependent ATPases including MCM proteins involved in the initiation of eukaryotic DNA replication. *Nucleic acids research*, 21(11), pp.2541–7.

Korhonen, J. a et al., 2004. Reconstitution of a minimal mtDNA replisome in vitro. *The EMBO journal*, 23(12), pp.2423–9.

Korhonen, J. a, Gaspari, M. & Falkenberg, M., 2003. TWINKLE Has 5' -> 3' DNA helicase activity and is specifically stimulated by mitochondrial single-stranded DNA-binding protein. *The Journal of biological chemistry*, 278(49), pp.48627–32.

Krude, T. & Musahl, C., 1996. Human replication proteins hCdc21, hCdc46 and P1Mcm3 bind chromatin uniformly before S-phase and are displaced locally during DNA replication. *Journal of cell science*, 318(Pt 2), pp.309–318.

Krueger, S. et al., 2014. The solution structure of full-length dodecameric MCM by SANS and molecular modeling. *Proteins*, (April), pp.1–11.

Kuchta, R.D. & Stengel, G., 2010. Mechanism and evolution of DNA primases. *Biochimica et Biophysica Acta*, 1804(5), pp.1180–1189.

Kusakabe, T. & Richardson, C.C., 1996. The Role of the Zinc Motif in Sequence Recognition by DNA Primases. *Journal of Biological Chemistry*, 271(32), pp.19563–19570.

Kutter, S. et al., 2007. The Influence of Protein Concentration on Oligomer Structure and Catalytic Function of Two Pyruvate Decarboxylases. *protein J*, 26, pp.585–591.

Kuznetsova, I.M., Turoverov, K.K. & Uversky, V.N., 2014. What Macromolecular Crowding Can Do to a Protein. *International Journal of Molecular Scienceournal of Molecular Science*, 15(12), pp.23090–23140.

Kyriakouli, D.S. et al., 2008. Progress and prospects : gene therapy for mitochondrial DNA disease. *Gene Therapy*, 15(14), pp.1017–1023.

De la Rosa-Trevín, J.M. et al., 2013. Xmipp 3.0: an improved software suite for image processing in electron microscopy. *Journal of structural biology*, 184(2), pp.321–8.

Larkin, M. a et al., 2007. Clustal W and Clustal X version 2.0. *Bioinformatics (Oxford, England)*, 23(21), pp.2947–8.

Leal, G.F. et al., 2003. A novel locus for autosomal recessive primary microcephaly (MCPH6) maps to 13q12.2. *Journal of medical genetics*, 40(7), pp.540–2.

Lee, J.K. & Hurwitz, J., 2000. Isolation and characterization of various complexes of the minichromosome maintenance proteins of Schizosaccharomyces pombe. *The Journal of biological chemistry*, 275(25), pp.18871–8.

Lee, J.K. & Hurwitz, J., 2001. Processive DNA helicase activity of the minichromosome maintenance proteins 4, 6, and 7 complex requires forked DNA structures. *Proceedings of the National Academy of Sciences of the United States of America*, 98(1), pp.54–9.

Lee, S. et al., 2012. Zinc-binding Domain of the Bacteriophage T7 DNA Primase Modulates Binding to the DNA Template. *Journal of biochemistry*, 287(46), pp.39030–39040.

Lee, S. & Richardson, C.C., 2002. Interaction of adjacent primase domains within the hexameric gene 4 helicase-primase of bacteriophage T7. *Proceedings of the National Academy of Sciences*, 99(20), pp.12703–12708.

Li, S. et al., 2012. Three-dimensional structure of basal body triplet revealed by electron cryo-tomography. *The EMBO journal*, 31(3), pp.552–62.

Lin, Y.-C. et al., 2013. Human microcephaly protein CEP135 binds to hSAS-6 and CPAP, and is required for centriole assembly. *The EMBO journal*, 32(8), pp.1141–54.

Lin, Y.-N. et al., 2013. CEP120 interacts with CPAP and positively regulates centriole elongation. *The Journal of cell biology*, 202(2), pp.211–9.

Linding, R., 2003. GlobPlot: exploring protein sequences for globularity and disorder. *Nucleic Acids Research*, 31(13), pp.3701–3708.

Liu, C. et al., 2012. Structural insights into the Cdt1-mediated MCM2-7 chromatin loading. *Nucleic acids research*, 40(7), pp.3208–17.

Liu, W. et al., 2008. Structural analysis of the Sulfolobus solfataricus MCM protein N-terminal domain. *Nucleic acids research*, 36(10), pp.3235–43.

Longley, M.J. et al., 2010. Disease variants of the human mitochondrial DNA helicase encoded by C10orf2 differentially alter protein stability, nucleotide hydrolysis, and helicase activity. *The Journal of biological chemistry*, 285(39), pp.29690–702.

Ludtke, S.J., Baldwin, P.R. & Chiu, W., 1999. EMAN: semiautomated software for high-resolution single-particle reconstructions. *Journal of structural biology*, 128(1), pp.82–97.

Lupas, A., M, V.D. & Stock, J., 1991. Predicting coiled coils from protein sequences. *Science*, 252(5009), pp.1162–4.

Lyubimov, A.Y. et al., 2012. ATP-dependent conformational dynamics underlie the functional asymmetry of the replicative helicase from a minimalist eukaryote. *Proceedings of the National Academy of Sciences of the United States of America*, 109(30), pp.11999–2004.

Ma, X. et al., 2010. The effects of oligomerization on Saccharomyces cerevisiae Mcm4/6/7 function. *BMC biochemistry*, 11(37), pp.1–15.

Madine, M. et al., 1995. The nuclear envelope prevents reinitiation of replication by regulating the binding of MCM3 to chromatin in Xenopus egg extracts. *Current Biology*, 5(11), pp.1270–1279.

Mahrenholz, C.C. et al., 2011. Complex networks govern coiled-coil oligomerization--predicting and profiling by means of a machine learning approach. *Molecular & cellular proteomics : MCP*, 10(5), p.M110.004994.

Maine, G., Sinha, P. & Tye, B., 1984. Mutants of S. cerevisiae defective in the maintenance of minichromosomes. *Genetics*, (Hand 1978).

Manuscript, A., 2011. Definition and estimation of resolution in single-particle reconstructions. *Structure*, 18(7), pp.768–775.

Mao, C. & Holt, I.J., 2009. Clinical and Molecular Aspects of Diseases of Mitochondrial DNA Instability. *Chang Gung Med.*, 32, pp.354–369.

Marsh, J. a & Teichmann, S. a, 2014. Protein flexibility facilitates quaternary structure assembly and evolution. *PLoS biology*, 12(5), p.e1001870.

Marsh, J.A. & Teichmann, S.A., 2013. Parallel dynamics and evolution: Protein conformational fluctuations and assembly reflect evolutionary changes in sequence and structure. *Bioessays*, 36(2), pp.209–218.

Marsh, J.A. & Teichmann, S.A., 2011. Relative Solvent Accessible Surface Area Predicts Protein Conformational Changes upon Binding. *Structure/Folding and Design*, 19(6), pp.859–867.

Matson, S.W., Richardsonq, C.C. & A, A.S.U.S., 1985. Nucleotide-dependent Binding of the Gene 4 Protein of Bacteriophage T7 to Single-stranded DNA. *Journal of Biological Chemistry*, 260(4), pp.2281–7.

Matsushima, Y. & Kaguny S., L., 2009. Functional importance of the conserved N-terminal domain of the mitochondrial replicative helicase. *Biochimica et biophysica acta*, 1787(5), pp.290–295.

Mcbride, H.M. & Neuspiel, M., 2006. Mitochondria : More Than Just a Powerhouse. *Current Biology*, 16(14), pp.551–560.

McGeoch, A.T. et al., 2005. Organization of the archaeal MCM complex on DNA and implications for the helicase mechanism. *Nature structural & molecular biology*, 12(9), pp.756–62.

Mcm, T., 1994. The MCM2-3-5 proteins : are they replication licensing factors ? *Trends in Cell Biology*, 4(5), pp.160–166.

Milazzo, A. et al., 2011. Characterization of a Direct Detection Device Imaging Camera for Transmission Electron Microscopy. *Ultramicroscopy*, 110(7), pp.744–747.

Miller, J.M. et al., 2014. Analysis of the crystal structure of an active MCM hexamer. *eLife*, 3, pp.1–13.

Ming, D. et al., 2002. How to describe protein motion without amino acid sequence and atomic coordinates. *Proceedings of the National Academy of Sciences*, 99(13), pp.8620–5.

Moir, D.O.N. & Botstein, D., 1982. COLD-SENSITIVE CELL-DIVISION-CYCLE MUTANTS OF YEAST: ISOLATION, PROPERTIES, AND PSEUDOREVERSION STUDIES. *Genetics*, 100(4), pp.547–563.

Murphy, E. & A. Eisner, D., 2010. Regulation of Intracellular and Mitochondrial Sodium in Health and Disease. *Circulation Research*, 104(3), pp.292–303.

Musahl, C. et al., 1995. A Human Homologue of the Yeast Replication Protein Cdc21. Interactions with other Mcm proteins. *European Journal of Biochemistry*, 230(3), pp.1096–1101.

Nguyen, T. et al., 2012. Interactions of the human MCM-BP protein with MCM complex components and Dbf4. *PloS one*, 7(4), p.e35931.

Nitani, N. et al., 2008. Mcm4 C-terminal domain of MCM helicase prevents excessive formation of single-stranded DNA at stalled replication forks. *Proceedings of the National Academy of Sciences of the United States of America*, 105(35), pp.12973–8.

Nogales-Cadenas, R. et al., 2009. CentrosomeDB: a human centrosomal proteins database. *Nucleic acids research*, 37(Database issue), pp.D175–80.

Nogueira, C. et al., 2014. Syndromes associated with mitochondrial DNA depletion. *Italian Journal of Pediatriics*, 40(34), pp.1–10.

Noree, C. et al., 2010. Identification of novel filament-forming proteins in. *Journal of Cell Biology*, 190(4), pp.541–551.

Notarnicola, S. et al., 1995. A domain of the gene 4 helicase/primase of bacteriophage T7 required for the formation of an active hexamer. *Journal of Biological Chemistry*, 270(34), pp.20215–24.

O'Shea, V.L. and Berger, J.M., 2014. Loading strategies of ring-shaped nucleic acid translocases and helicases. *Current opinion in structural biology*, 25, pp.16–24.

Pascual-Montano, A. et al., 2006. Optimization problems in electron microscopy of single particles. *Ann. Oper. Res.*, 148, pp.133–165.

Patel, S.S. & Picha, K.M., 2000. Structure and function of hexameric helicases. *Annu. Rev. Biochem.*, 3(69), pp.651–97.

Patelt, S.S., 1995. Bacteriophage T7 helicase / primase proteins form rings around single-stranded DNA that suggest a general structure for hexameric helicases. *Proceedings of the National Academy of Sciences*, 92(9), pp.3869–3873.

Petrovska, I. et al., 2014. Filament formation by metabolic enzymes is a specific adaptation to an advanced state of cellular starvation. *eLife*, (April), pp.1–19.

Pettersen, E.F. et al., 2004. UCSF Chimera--a visualization system for exploratory research and analysis. *Journal of computational chemistry*, 25(13), pp.1605–12.

Pham, X.H. et al., 2006. Conserved Sequence Box II Directs Transcription Termination and Primer Formation in Mitochondria. *Journal of Biological Chemistry*, 281(34), pp.24647–24652.

Podobnik, M. et al., 2000. A TOPRIM Domain in the Crystal Structure of the Catalytic Core of Escherichia coli Primase Confirms a Structural Link to DNA Topoisomerases. *Journal of molecular biology*, 300(2), pp.353–362.

Prokhorova, T.A. & Blow, J.J., 2000. Sequential MCM / P1 Subcomplex Assembly Is Required to Form a Heterohexamer with Sequential MCM / P1 Subcomplex Assembly Is Required to Form a Heterohexamer with Replication Licensing Activity *. *Journal of Biological Chemistry*, 275(4), pp.2491–2498.

Pucci, B. et al., 2007. Modular organization of the Sulfolobus solfataricus mini-chromosome maintenance protein. *The Journal of biological chemistry*, 282(17), pp.12574–82.

Radermacher, M., 1988. Three dimensional reconstruction of single particles from random and nonrandom tilt series . *Journal of Electron Microscopy Technique*, 9(4), p.1988.

Remus, D. et al., 2009. Concerted loading of Mcm2-7 double hexamers around DNA during DNA replication origin licensing. *Cell*, 139(4), pp.719–30.

Remus, D. & Diffley, J.F.X., 2009. Eukaryotic DNA replication control: lock and load, then fire. *Current opinion in cell biology*, 21(6), pp.771–7.

Reyes, A. et al., 2013. Mitochondrial DNA replication proceeds via a " bootlace " mechanism involving the incorporation of processed transcripts. *Nucleic acids research*, 41(11), pp.5837–5850.

Richter, A. & Knippers, R., 1997. High-molecular-mass complexes of human minichromosome-maintenance proteins in mitotic cells. *European Journal of Biochemistry*, 141, pp.136–141.

Robinson, C. V, Sali, A. & Baumeister, W., 2007. The molecular sociology of the cell. *Nature*, 450(7172), pp.973–982.

Romero, P. & Obradovic, Z., 1998. Thousands of proteins likely to have long disordered regions. *Pac. Symp. Biocomput*, pp.437–48.

Rosenthal, P.B. & Henderson, R., 2003. Optimal Determination of Particle Orientation, Absolute Hand, and Contrast Loss in Single-particle Electron Cryomicroscopy. *Journal of Molecular Biology*, 333(4), pp.721–745.

Rossmann, M.G. et al., 2005. Combining X-ray crystallography and electron microscopy. *Structure (London, England : 1993)*, 13(3), pp.355–62.

Rothenberg, E. et al., 2007. MCM forked substrate specificity involves dynamic interaction with the 5′-tail. *Journal of Biological Chemistry*, 282(47), pp.34229–34234.

Roy, A., Kucukural, A. & Zhang, Y., 2010. I-TASSER: a unified platform for automated protein structure and function prediction. *Nature protocols*, 5(4), pp.725–38.

Sakakibara, N., Kelman, L.M. & Kelman, Z., 2009. Unwinding the structure and function of the archaeal MCM helicase. *Molecular Microbiology*, 72(2), pp.286–296.

Dos Santos, H.G. et al., 2013. Structure and non-structure of centrosomal proteins. *PloS one*, 8(5), p.e62633.

Sato, M. et al., 2000. Electron microscopic observation and single-stranded DNA binding activity of the Mcm4,6,7 complex. *Journal of molecular biology*, 300(3), pp.421–31.

Scheffler, I.E., 2007. *Mitochondria* 2nd ed. John Wiley & Sons, ed., Hoboken.

Scheres, S.H.W., 2012. RELION: implementation of a Bayesian approach to cryo-EM structure determination. *Journal of structural biology*, 180(3), pp.519–30.

Scheres, S.H.W. & Chen, S., 2012. Prevention of overfitting in cryo-EM structure determination The Coherent X-ray Imaging Data Bank. *Nature Methods*, 9(9), pp.853–854.

Schroer, M.A. et al., 2012. The Effect of Ionic Strength , Temperature , and Pressure on the Interaction Potential of Dense Protein Solutions : From Nonlinear Pressure Response to Protein Crystallization. , 102(June), pp.2641–2648.

Schulte, D., Burkhart, R. & Musahl, C., 1995. Expression, phosphorylation and nuclear localization of the human P1 protein, a homologue of the yeast Mcm 3 replication protein. *Journal of cell science*, 108(Pt 4), pp.1381–1389.

Schwacha, A. & Bell, S.P., 2001. Interactions between Two Catalytically Distinct MCM Subgroups Are Essential for Coordinated ATP Hydrolysis and DNA Replication. *Molecular Cell*, 8(5), pp.1093–1104.

Sen, D. et al., 2012. Human mitochondrial DNA helicase TWINKLE is both an unwinding and annealing helicase. *The Journal of biological chemistry*, 287(18), pp.14545–56.

Sherman, D. & Forsburg, S., 1998. Schizosaccharomyces pombe Mcm3p, an essential nuclear protein, associates tightly with Nda4p (Mcm5p). *Nucleic acids research*, 26(17), pp.3955–3960.

Sherman, D.A., Pasion, S.G. & Forsburg, S.L., 1998. Multiple domains of fission yeast Cdc19p (MCM2) are required for its association with the core MCM complex. *Molecular Biology of the Cell*, 9(7), pp.1833–1845.

Sheu, Y.-J. et al., 2014. Domain within the helicase subunit Mcm4 integrates multiple kinase signals to control DNA replication initiation and fork progression. *Proceedings of the National Academy of Sciences of the United States of America*, 111(18), pp.E1899–908.

Shi, Y. et al., 2008. Human mitochondrial RNA polymerase primes lagging-strand DNA synthesis in vitro. *Proceedings of the National Academy of Sciences*, 105(32), pp.11122–7.

Shutt, T.E. & Gray, M.W., 2006a. Bacteriophage origins of mitochondrial replication and transcription proteins. *Trends in genetics*, 22(2), pp.90–95.

Shutt, T.E. & Gray, M.W., 2006b. Twinkle, the mitochondrial replicative DNA helicase, is widespread in the eukaryotic radiation and may also be the mitochondrial DNA primase in most eukaryotes. *Journal of molecular evolution*, 62(5), pp.588–99.

Singleton, M.R. et al., 2000. Crystal Structure of T7 Gene 4 Ring Helicase Indicates a Mechanism for Sequential Hydrolysis of Nucleotides. *Cell*, 101, pp.589–600.

Singleton, M.R., Dillingham, M.S. & Wigley, D.B., 2007. Structure and mechanism of helicases and nucleic acid translocases. *Annual review of biochemistry*, 76, pp.23–50.

Sorzano, C.O.S. et al., 2010. A clustering approach to multireference alignment of single-particle projections in electron microscopy. *Journal of structural biology*, 171(2), pp.197–206.

Spelbrink, J.N. et al., 2001. Human mitochondrial DNA deletions associated with mutations in the gene encoding Twinkle, a phage T7 gene 4-like protein localized in mitochondria. *Nature genetics*, 28(3), pp.223–31.

Sterner, J.M. et al., 1998. Negative Regulation of DNA Replication by the Retinoblastoma Protein Is Mediated by Its Association with MCM7 Negative Regulation of DNA Replication by the Retinoblastoma Protein Is Mediated by Its Association with MCM7. *Molecular and cellular biology*, 18(5), pp.2748–2757.

Su, T.T., Feger, G. & Farrell, P.H.O., 1996. Drosophila MCM Protein Complexes. *Molecular Biology of the Cell*, 7(2), pp.319–329.

Sun, J. et al., 2013. Cryo-EM structure of a helicase loading intermediate containing ORC-Cdc6-Cdt1-MCM2-7 bound to DNA. *Nature structural & molecular biology*, 20(8), pp.944–51.

Sun, J. et al., 2014. Structural and mechanistic insights into Mcm2 – 7 double-hexamer assembly and function. *Genes & development*, 28(20), pp.2291–2303.

Suomalainen, A. et al., 1997. Autosomal dominant progressive external ophthalmoplegia with multiple deletions of mtDNA: clinical, biochemical, and molecular genetic features of the 10q-linked disease. *Neurology*, 48(5), pp.1244–53.

Takahashi, T.S., Wigley, D.B. & Walter, J.C., 2005. Pumps, paradoxes and ploughshares: mechanism of the MCM2-7 DNA helicase. *Trends in biochemical sciences*, 30(8), pp.437–44.

Tama, F., Wriggers, W. & Chaco, P., 2003. Mega-Dalton Biomolecular Motion Captured from Electron Microscopy Reconstructions. *Journal of molecular biology*, 2836(02), pp.485–492.

Tama, F., Wriggers, W. & Iii, C.L.B., 2002. Exploring Global Distortions of Biological Macromolecules and Assemblies from Low-resolution Structural Information and Elastic Network Theory. *Journal of molecular biology*, 2836(02), pp.297–305.

Tang, C.-J.C. et al., 2009. CPAP is a cell-cycle regulated protein that controls centriole length. *Nature Cell Biology*, 11(7), pp.825–831.

Tang, C.-J.C. et al., 2011. The human microcephaly protein STIL interacts with CPAP and is required for procentriole formation. *The EMBO Journal*, 30(23), pp.4790–804.

Tompa, P. & Fuxreiter, M., 2008. Fuzzy complexes: polymorphism and structural disorder in protein-protein interactions. *Trends in biochemical sciences*, 33(1), pp.2–8.

Toth, E. a. et al., 2003. The crystal structure of the bifunctional primase-helicase of bacteriophage T7. *Molecular Cell*, 12(5), pp.1113–1123.

Tougu, K., Peng, H. & Marians, J., 1994. Identification of a Domain of Escherichia coli Primase Required for Functional Interaction with the DnaB Helicase at the Replication Fork. *Journal of Biological Chemistry*, 269(6), pp.4675–4682.

Tower, J., 2015. Programmed cell death in aging. *Ageing Research Reviews*, (pii: S1568-1637(15)00036-7), pp.1–11.

Tran, N.Q. et al., 2010. A single subunit MCM6 from pea forms homohexamer and functions as DNA helicase. *Plant molecular biology*, 74(4-5), pp.327–36.

Treviño, M. a et al., 2014. Emergence of structure through protein-protein interactions and pH changes in dually predicted coiled-coil and disordered regions of centrosomal proteins. *Biochimica et biophysica acta*.

Tsuruga, H. et al., 1997. Expression, nuclear localization and interactions of human MCM/P1 proteins. *Biochemical and biophysical research communications*, 236(1), pp.118–25.

Tye, B.I.K.K., 1996. Physical Interactions among Mcm Proteins and Effects of Mcm Dosage on DNA Replication in Saccharomyces cerevisiae. *Molecular and cellular biology*, 16(9), pp.5081–5090.

Tye, B.K. & Sawyer, S., 2000. The hexameric eukaryotic MCM helicase: building symmetry from nonidentical parts. *The Journal of biological chemistry*, 275(45), pp.34833–6.

Tyynismaa, H. et al., 2004. Twinkle helicase is essential for mtDNA maintenance and regulates mtDNA copy number. *Human molecular genetics*, 13(24), pp.3219–3227.

Uhn, L.E.A.K., 1999. The accessory subunit of mtDNA polymerase shares structural homology with aminoacyl-tRNA synthetases : Implications for a dual role as a primer recognition factor and processivity clamp. *Proceedings of the National Academy of Sciences*, 96(17), pp.9527–9532.

Vargas, J. et al., 2014. Efficient initial volume determination from electron microscopy images of single particles. *Bioinformatics (Oxford, England)*, pp.1–8.

Vincent, T.L., Green, P.J. & Woolfson, D.N., 2013. LOGICOIL--multi-state prediction of coiled-coil oligomeric state. *Bioinformatics (Oxford, England)*, 29(1), pp.69–76.

Vulprecht, J. et al., 2012. STIL is required for centriole duplication in human cells. *Journal of cell science*, 125(Pt 5), pp.1353–62.

Wanrooij, S. et al., 2010. Mitochondrial RNA Polymerase Is Needed for Activation of the Origin of Light-Strand DNA Replication. *Molecular Cell*, 37(1), pp.67–78.

Wanrooij, S. & Falkenberg, M., 2010. The human mitochondrial replication fork in health and disease. *BBA - Bioenergetics*, 1797(8), pp.1378–1388.

Watanabe, E., Ohara, R. & Ishimi, Y., 2012. Effect of an MCM4 mutation that causes tumours in mouse on human MCM4/6/7 complex formation. *Journal of biochemistry*, 152(2), pp.191–8.

Wei, Z. et al., 2010. Characterization and structure determination of the Cdt1 binding domain of human minichromosome maintenance (Mcm) 6. *The Journal of biological chemistry*, 285(17), pp.12469–73.

Woodward, A.M. et al., 2006. Excess Mcm2-7 license dormant origins of replication that can be used under conditions of replicative stress. *The Journal of cell biology*, 173(5), pp.673–83.

Wu, Y., 2012. Unwinding and rewinding: double faces of helicase? *Journal of nucleic acids*, 2012, p.140601.

Xu, M., Chang, Y.P. & Chen, X.S., 2013. Expression, purification and biochemical characterization of Schizosaccharomyces pombe Mcm4, 6 and 7. *BMC biochemistry*, 14, p.5.

Yabuta, N. et al., 2003. Mammalian Mcm2/4/6/7 complex forms a toroidal structure. *Genes to Cells*, 8(5), pp.413–421.

Yakovchuk, P., Protozanova, E. & Frank-Kamenetskii, M.D., 2006. Base-stacking and base-pairing contributions into thermal stability of the DNA double helix. *Nucleic acids research*, 34(2), pp.564–74.

Yankulov, K. et al., 1999. MCM Proteins Are Associated with RNA Polymerase II Holoenzyme MCM Proteins Are Associated with RNA Polymerase II Holoenzyme. *Molecular and cellular biology*, 19(9).

Ying, C.Y. & Gautier, J., 2005. The ATPase activity of MCM2-7 is dispensable for pre-RC assembly but is required for DNA unwinding. *The EMBO journal*, 24(24), pp.4334–44.

You, Z. et al., 2002. Roles of Mcm7 and Mcm4 subunits in the DNA helicase activity of the mouse Mcm4/6/7 complex. *The Journal of biological chemistry*, 277(45), pp.42471–9.

You, Z., Komamura, Y. & Ishimi, Y., 1999. Biochemical analysis of the intrinsic Mcm4-Mcm6-mcm7 DNA helicase activity. *Molecular and cellular biology*, 19(12), pp.8003–15.

You, Z. & Masai, H., 2008. Cdt1 forms a complex with the minichromosome maintenance protein (MCM) and activates its helicase activity. *The Journal of biological chemistry*, 283(36), pp.24469–77.

You, Z. & Masai, H., 2005. DNA binding and helicase actions of mouse MCM4/6/7 helicase. *Nucleic acids research*, 33(9), pp.3033–47.

Yu, X. et al., 2002. The Methanobacterium thermoautotrophicum MCM protein can form heptameric rings. *EMBO reports*, 3(8), pp.792–797.

Yu, Z., Feng, D. & Liang, C., 2004. Pairwise interactions of the six human MCM protein subunits. *Journal of molecular biology*, 340(5), pp.1197–206.

Zhang, L. et al., 2013. Structural basis of transfer between lipoproteins by cholesteryl ester transfer protein. *Nature Chemical Biology*, 8(4), pp.342–349.

Zhang, Y., 2008. I-TASSER server for protein 3D structure prediction. *BMC bioinformatics*, 9, p.40.

Zhao, H. et al., 2013. The Cep63 paralogue Deup1 enables massive de novo centriole biogenesis for vertebrate multiciliogenesis. *Nature cell biology*, 15(12), pp.1434–44.

Zhao, L. et al., 2010. Dimerization of CPAP Orchestrates Centrosome Cohesion Plasticity. *The Journal of Biological Chemistry*, 285(4), pp.2488–2497.

Zheng, T. et al., 2014. Plasma minichromosome maintenance complex component 6 is a novel biomarker for hepatocellular carcinoma patients. *Hepatology research*, pp.1–10.

Zheng, X. et al., 2014. Conserved TCP domain of Sas-4/CPAP is essential for pericentriolar material tethering during centrosome biogenesis. *Proceedings of the National Academy of Sciences of the United States of America*, 111(3), pp.E354–63.

Zhu, J. & Frank, J., 1997. Three-Dimensional Reconstruction with Contrast Transfer Function Correction from Energy-Filtered Cryoelectron Micrographs : Procedure and Application to the 70S Escherichia coli Ribosome. *Journal of structural biology*, 118(3), pp.197–219.

Ziebarth, T.D. et al., 2010. Dynamic effects of cofactors and DNA on the oligomeric state of human mitochondrial DNA helicase. *The Journal of biological chemistry*, 285(19), pp.14639–47.

Ziebarth, T.D., Farr, C.L. & Kaguni, L.S., 2007. Modular architecture of the hexameric human mitochondrial DNA helicase. *Journal of molecular biology*, 367(5), pp.1382–91.

Apendix

Publications resulting from projects apart from the ones included in the present thesis:

Sorzano CO, Vargas J, de la Rosa-Trevin JM, Oton J, Alvarez-Cabrera AL, Abrishami V, Sesmero E, Marabini R, Carazo JM. 2015. A statistical approach to the initial volume problem in Single Particle Analysis by Electron Microscopy. Journal of Structural Biology, 189 (3), pp 213-219.

Vargas J, Álvarez-Cabrera AL, Marabini R, Carazo JM, Sorzano CO. 2014. Efficient initial volume determination from electron microscopy images of single particles.. Bioinformatics, 30(20), pp 2891-2898.

S. Jonic, C.O.S. Sorzano, A.L. Álvarez-Cabrera, J.M. Carazo. StructMap: Multivariate distance analysis of elastically aligned electron microscopy structures for exploring pathways of conformational changes. 2015. *Structure*. Paper under revision.

BIOINFORMATICS ORIGINAL PAPER

Vol. 30 no. 20 2014, pages 2891–2898
doi:10.1093/bioinformatics/btu404

Structural bioinformatics

Advance Access publication June 27, 2014

Efficient initial volume determination from electron microscopy images of single particles

Javier Vargas[1,*], Ana-Lucia Álvarez-Cabrera[1], Roberto Marabini[2], Jose M. Carazo[1] and C. O. S. Sorzano[1]

[1]Biocomputing Unit, Centro Nacional de Biotecnología-CSIC, C/Darwin 3 and [2]Escuela Politécnica Superior, Universidad Autónoma de Madrid, C/Francisco Tomás y Valiente, 28049, Cantoblanco (Madrid), Spain

Associate Editor: Robert Murphy

ABSTRACT

Motivation: Structural information of macromolecular complexes provides key insights into the way they carry out their biological functions. The reconstruction process leading to the final 3D map requires an approximate initial model. Generation of an initial model is still an open and challenging problem in single-particle analysis.

Results: We present a fast and efficient approach to obtain a reliable, low-resolution estimation of the 3D structure of a macromolecule, without any a priori knowledge, addressing the well-known issue of initial volume estimation in the field of single-particle analysis. The input of the algorithm is a set of class average images obtained from individual projections of a biological object at random and unknown orientations by transmission electron microscopy micrographs. The proposed method is based on an initial non-lineal dimensionality reduction approach, which allows to automatically selecting representative small sets of class average images capturing the most of the structural information of the particle under study. These reduced sets are then used to generate volumes from random orientation assignments. The best volume is determined from these guesses using a random sample consensus (RANSAC) approach. We have tested our proposed algorithm, which we will term 3D-RANSAC, with simulated and experimental data, obtaining satisfactory results under the low signal-to-noise conditions typical of cryo-electron microscopy.

Availability: The algorithm is freely available as part of the Xmipp 3.1 package [http://xmipp.cnb.csic.es].

Contact: jvargas@cnb.csic.es

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on February 15, 2014; revised and accepted on June 23, 2014

1 INTRODUCTION

Single-Particle Analysis (SPA) techniques can obtain 3D maps of biological complexes at near-atomic resolution by combining tens of thousands of projection images obtained with a transmission electron microscopy (TEM) (Frank, 1996; Zhang and Zhou, 2011). In general, the reconstruction process leading to the final 3D map requires the use of an approximate initial model. The fully automatic and efficient determination of this initial volume, either for symmetric or asymmetric structures, is still an open and challenging problem in SPA, as indicated by the existence of an ample literature on this matter.

Many previous attempts to the 'initial model problem' have been reported. On the one hand, in the Random Conical Tilt and orthogonal tilt methods (Leschziner and Nogales, 2006; Radermacher *et al.*, 1987), the alignment problem is simplified by acquiring micrographs as tilt pairs, or by using multiple different tilts with known tilt angles. On the other hand, we observe a large variety of methods based on common lines (Castón *et al.*, 1999; Crowther *et al.*, 1970; Elmlund and Elmlund, 2012; Elmlund *et al.*, 2008, 2010; Liu *et al.*, 2007; Ogura and Sato, 2006; Penczek and Zhu, 1996; van Heel, 1987; Thuman-Commike and Chiu, 1997) that, in principle, they allow for an initial model estimation without tilting of the specimen. Computer-generated shapes (Baker and Cheng, 1996; Bilbao-Castro *et al.*, 2004; Ludtke *et al.*, 2004), or reconstructions from one image of a particle assuming a certain symmetry (Cantele *et al.*, 2003; Castón *et al.*, 1999), have also been used to define initial volumes. Another line of research is based on the introduction of a 'random model' strategy based on first assigning random orientations to class averages (Harauz and van Heel, 1985; van Heel, 1984). In this latter case, an initial 3D reconstruction is obtained from these random angular assignments, which is finally refined by a projection matching strategy. Following this approach, Sanz-Garcia *et al.*, 2010 presented a random-model method that allows *ab initio* generation of starting models from raw experimental images. Several initial models were generated, assigning initially a random orientation to each imaged particle. Recently, Elmlund *et al.*, 2013, have presented a method based on a probabilistic initial 3D model generation procedure, which uses projection images instead of class averages. Furthermore, in Lyumkis *et al.* (2013), it presented OptiMod, a method that incorporates multiple automated algorithms for determining orientations using common-lines methodologies and, at the same time, provides criteria for scoring their results. This approach generates multiple maps using algorithm-specific randomization.

The 'initial model problem' is still, and in spite of the multiple algorithms so far proposed, widely accepted as a real issue in SPA (Taylor and Glaeser, 2008, Voss *et al.*, 2010), with methods demanding no trivial choices of input parameters and being computationally expensive. Indeed, in the way of defining high-throughput approaches in 3D-EM, we really need fast, simple and accurate methods, which are, indeed, the motivation of this

*To whom correspondence should be addressed.

work. In this way, we will describe in detail 3D-RANSAC, presenting its good performance on a wide range of specimens, using the same parameters for all cases (making the case of 'simplicity'), and obtaining final results in a matter of minutes on a typical laptop computer. Our proposed approach consists in a novel random modelling strategy based on an initial dimensionality-reduction method together with the RANSAC algorithm, which makes this approach efficient from a computational point of view. The method can be used to produce low-resolution initial volumes of symmetric or asymmetric biological complexes.

2 METHODS

2.1 Class average images

First, a set of class average images are obtained from the particle dataset using any image classification algorithm [in this article, we will use CL2D (Sorzano *et al*., 2010), included in the Xmipp 3.1 package]. Observe that the classification process is required in any usual SPA processing workflow and is not a special requirement of the proposed approach. Electron microscopy datasets commonly contain more than one different structures, as projections of different conformations of a given molecule, or projections of different molecules in the specimen preparation. These class images are then the input of the new proposed method and, typically, ~20–50 class images are sufficient. There is nothing in the algorithm that precludes using experimental projections instead of classes. However, we find that using a sufficient number of class average images has several practical advantages, such as (i) increasing the signal-to-noise ratio of the input images, (ii) introducing a desired smoothing (a '*de facto*' regularization) in the landscape of solutions and (iii) reducing the total processing time. Of course, we note that we are only interested in a low-resolution initial model.

2.2 Random model generation strategy

Our random model generation strategy is based on the following eight steps:

(1) The class averages are low-pass filtered, and their size is reduced according to user parameters (basically, the desired resolution in the initial model).

(2) The local tangent space alignment (LTSA) non-linear dimensionality reduction approach (Zhang and Hongyuan, 2005) is applied to automatically select a random, smaller and appropriate (in the sense that contains the most of the structural information of the biological complex under study) subset of class images. This approach non-linearly projects the class averages onto a lower-dimensional space [in our case, two-dimensional (2D)], where the projections of the structure at similar orientations appear close. LTSA method is essentially a local principal component analysis (PCA) approach that can efficiently 'learn' about non-linear manifolds by taking into account that non-linear manifolds can be considered to be locally linear in small neighborhoods. Observe that PCA cannot deal with non-linear manifolds, as it is a linear method, and the 2D projection of a 3D structure is, intrinsically, a non-linear process (Giannakis *et al*, 2012). Additionally, we note that the computer time required for LTSA and PCA are comparable, making LTSA clearly the method of choice for this task. As an example, in Figure 1, we show a projection of a set of image class averages into a 2D space. As can be seen from Figure 1, and as intuitively expected, similar projections of the biological object appear close in the 2D space, whereas different ones appear far apart. We can now use this 2D space to select a smaller image dataset containing the most of the structural information, and then assign random orientations to these images. To automatically select a representative image set, the 2D map is partitioned by a 2D regular grid of dimensions (with typically $n = 9$), and only one image class average is randomly selected each time from any of the grid squares. In Figure 1, we show an example of a smaller image dataset with a red square, and the 2D regular grid appears in gray color. Therefore, n class images are randomly selected in this way. Obviously, there is a trade-off with respect to the number of images composing this reduced dataset. As the number of images gets smaller, the processing time of subsequent steps is reduced and, additionally, the probability of assigning correct angles by random assignment is higher. However, if this number is too small, we can lose important structural information. In our algorithm, this number (n) is an input parameter and normally ranges between 4 and 9, with 9 being the 'default' value (all results presented in this article have been obtained with $n = 9$).

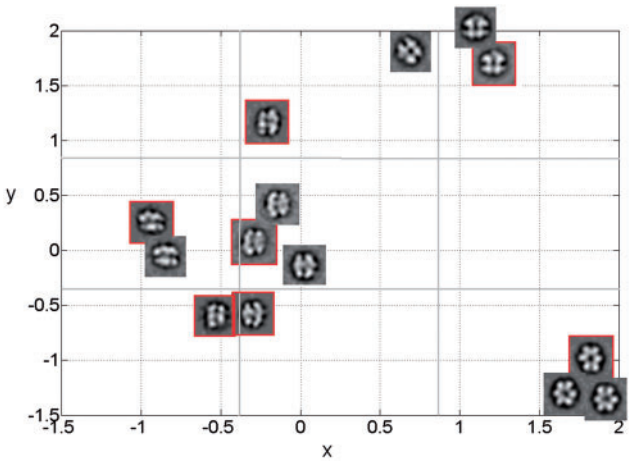


Fig. 1. Projection of a set of unsorted image class averages of a certain MCM467 complex into a 2D space using the LTSA dimensionality reduction approach. In this figure, x and y are the axis of the feature space learned from the input

(3) A 3D reconstruction is performed from the smaller image subset ($n = 9$) by random angular assignment. The reconstruction algorithm takes advantage of symmetry information when available. Reconstruction is made by interpolation in Fourier space. For each experimental image, the Fourier Transform is computed and placed in the corresponding plane in the 3D space as well as in all planes related by symmetry. This random 3D model is then projected at regular angular intervals determined by the angular step-size, which is an input parameter with a default value of 7°. Each initial class average is now compared with the projections of the 3D random model, and the best assignment is determined looking for the largest correlation coefficient.

(4) Steps 2–3 are repeated N times producing N different 3D random models. Therefore, N is also an input parameter of the method, but in the Supplementary Material, we present a statistical derivation for an informed selection of N. In this way, we set the default value of N to 380, as in this way, we know that the probability of obtaining a still better 3D model by increasing N is <0.01. This value of $N = 380$ has been used in all the examples presented in this article.

(5) For each generated random model, we define its inliers as the initial class average images that have a large enough correlation coefficient with respect to the reprojections of the random 3D map.

We will refer to these 'good' initial classes as 'inliers', and we say that they support this 3D model. The rest of the initial classes are referred to as 'outliers'. The practical way in which this selection is done is by establishing a threshold in the correlation coefficient obtained as the percentile p of all obtained correlations, with p typically between 75 and 80%. We define the score of each random 3D map as the sum of the correlations of its inliers. In steps 2–5, we are using Random Sample Consensus (RANSAC) approach (Fischler and Bolles, 1981) to capture 'correct' models. RANSAC is an iterative method to estimate a model (in this case a 3D map) from a set of observed data that contains a large amount of outliers. RANSAC approach consists in the following steps: (i) Randomly selecting a subset of the dataset. (ii) Fitting a model to the selected subset. (iii) Determining a score to each model, typically the number of outliers/inliers. (iv) Repeating steps 1–3 for a prescribed number of iterations. A basic assumption is that the data consist of inliers, data that can be explained by the model, though they may be subject to noise, and outliers, that are data that do not fit the model. The outliers may come, e.g. from extreme values of the noise, from erroneous measurements or incorrect hypotheses about the interpretation of data, for instance, wrongly assigned Euler angles. RANSAC also assumes that, given a (usually small) set of inliers, there exists a procedure that can estimate the model that optimally explains or fits this data. Additionally, this algorithm is probabilistic and non-deterministic, in the sense, that it produces always good results with a certain probability if the number of inliers is larger than the number of outliers, with this probability increasing as more iterations are allowed (Fischler and Bolles, 1981). Our RANSAC approach is performed using the following combination of steps: (i) Automatic selection of a small number of class averages following a dimensionality reduction approach. (ii) Random assignment of angles to each of the selected class averages and computation of a 3D model using them. (iii) Calculation of a score for each model as the sum of inliers correlation. (iv) Repeat steps 1–3 for a prescribed number of iterations. In the Supplementary Material, we show that with the number of RANSAC iterations >380, the probability of finding a better model is <0.01.

(6) The k random 3D models ($k \sim 5$–10) with largest number of inliers (largest score) are selected and new 3D reconstructions are obtained using as input classes only the inliers.

(7) The previous k 3D random models are now refined against all initial classes through a model refining strategy. In this article, we have used a projection matching approach (de la Rosa-Trevín *et al.*, 2013), using typically 10 iterations for refinement. However, other approaches can be used as well, such as the recent method of Elmlund *et al.*, 2013. Observe that we refine the best K models independently through projection matching to add more robustness to our approach. As the previous angular assignment is random, refining K models improves the probability of getting at least some good structures at the end.

(8) The resulting k volumes are scored taking into account the sum of the inliers correlation coefficients. Finally, the model with highest score is automatically selected.

In Figure 2, we show a diagram of the different processing steps.

3 RESULTS

In this section, we provide results obtained with simulated and experimental data that show the effectiveness of the proposed method.

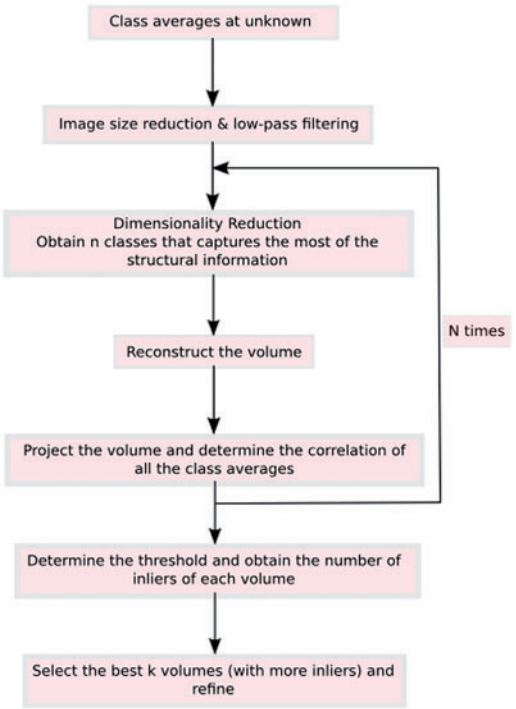


Fig. 2. Diagram of the different processing steps the input

3.1 Simulations

In the first simulation, we used the structure of the Bacteriorhodopsin as a phantom (PDB entry: 1BRD, Henderson *et al.*, 1990) that is projected at 200 unknown and random orientations. The projections are affected by a Gaussian noise with a signal-to-noise ratio (SNR) of 0.1. The size of the images is 100×100 px and the sampling rate is 1 Å/px. Note that, in this case, we did not obtain class averages and we use our proposed method directly with the noisy projections to show its robustness.

In Figure 3, we present 7 of the 200 noisy projections used, together with the phantom (left) and best obtained initial (right) 3D maps at three different orientations. The initial volume was obtained with the following parameters, $N = 380$, $n = 9$, $P = 0.77$ and $k = 10$. We used an angular sampling rate for retroprojection of 7°. The projections were low-pass filtered to a resolution of 5 Å. The required processing time for obtaining the $k = 10$ volumes is of 35 min with a 2.5 GHz laptop and using two processors.

To quantize the resolution of the obtained initial volumes, we have obtained the Fourier Shell Correlation (FSC) curves using the PDB volume as reference. The resolution at FSC = 0.5 and FSC = 0.143 are 4.6 and 4.5 Å, respectively (Fig. 4), that is consistent with the previously performed low-pass filtering, and shows that using perfectly aligned projections in presence of moderate noise, the proposed method can retrieve high-resolution models. This situation is not usual in experimental cases. However, with this simulation, we want to show that there is no theoretical restriction that limits the proposed method in high-resolution analysis.

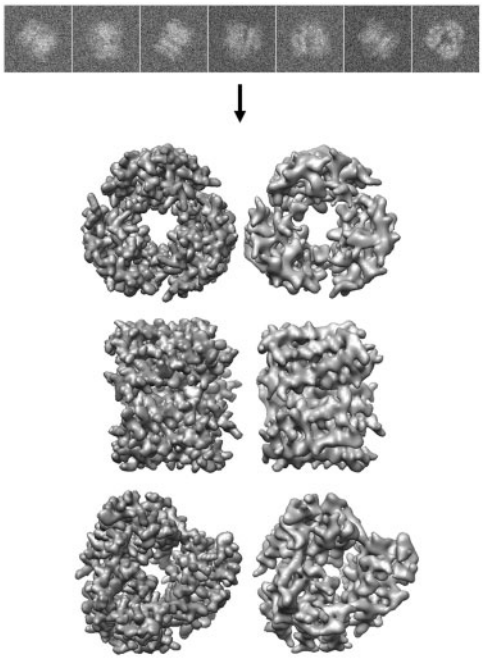


Fig. 3. Seven projections of Bacteriorhodopsin (PDB entry: 1BRD, Henderson *et al.*, 1990) phantom map with SNR of 0.1 are shown at top of the figure. The phantom 3D map at three different orientations is presented on the left-hand side, whereas the best obtained volume, using the proposed approach, is on the right-hand side

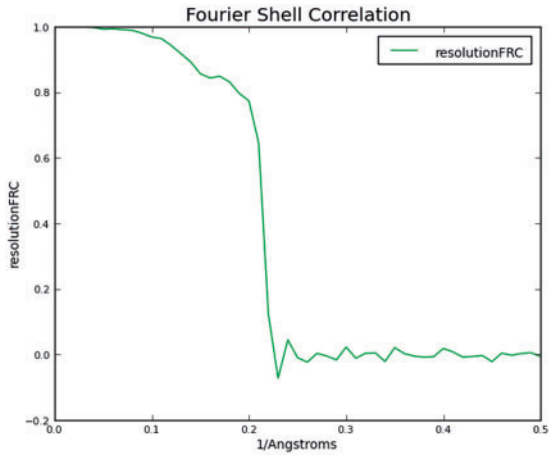


Fig. 4. FSC curve between the Bacteriorhodopsin phantom of the best initial 3D map obtained (highest score) using the proposed approach

3.2 Experimental results

3.2.1 Case 1: Bovine Papillomavirus The first case consists of images of Bovine Papillomavirus (BPV) (Wolf *et al.*, 2010) kindly made accessible by Drs Wolf and Grigorieff. The dataset consists of 49 micrographs of an approximate size of 10 000 × 10 000 pixels. The sampling rate is 1.237 Å/pixel, the microscope voltage 300 kV and the magnification × 56 588. A total of 5317 particles of size 120 × 120 px were identified using the method presented in (Abrishami *et al.*, 2013), from which 32 classes were determined

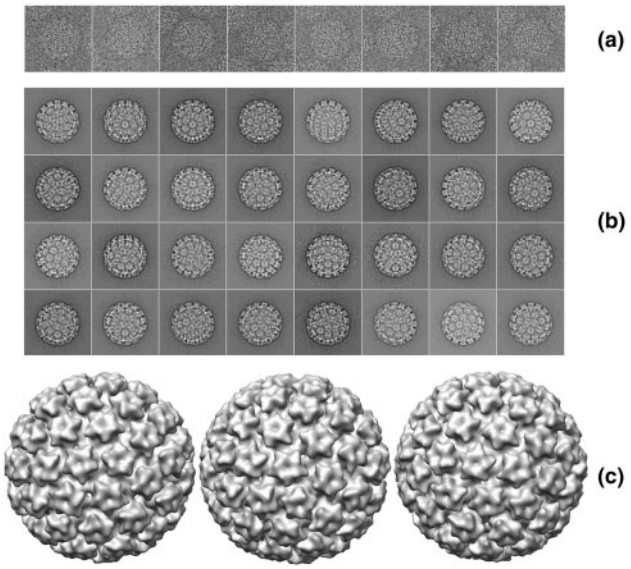


Fig. 5. Eight experimental projections of the BPV (**a**), 32 classes obtained using CL2D (**b**) and the best initial volume obtained by the proposed method at three different orientations (**c**)

using CL2D (Sorzano *et al.*, 2010). These classes were low-pass filtered to a resolution of 5 Å. In Figure 5, we show eight experimental projections (a) and the 32 obtained classes (b). The parameters used to obtain the initial volume by the proposed method are the same ones as in the case presented before. In Figure 5c, we show the best 3D-RANSAC map at three different orientations. The processing time for obtaining the k = 10 volumes is of ~20 min, with the same computer as before using two processors. Observe that the processing time of CL2D is of 330 min, and then the total time CL2D + 3D-RANSAC is of 350 min.

3.2.2 Case 2: Eukaryotic ribosome Moreover, we have also performed an asymmetric reconstruction using ~5000 cryo-EM projections of an eukaryotic ribosome, obtained from the EMDB test image data (http://www.ebi.ac.uk/pdbe/emdb/test_data.html) and originally used in the work of Scheres *et al.* (2007). The images have a size of 130 × 130 px. We initially obtained 16 class average images using CL2D, which were then low-pass filtered to a resolution of 25 Å. We obtained the 3D-RANSAC map with the same parameters as in the cases before. In Figure 6, we show eight initial experimental projections of the ribosome (a) and the low-pass filtered class averages (b). Finally, in Figure 6c, we also show the 3D reconstruction using PRIME (Elmlund *et al.*, 2013) and one obtained using 3D-RANSAC at three different orientations. The FSC curves between the PRIME and the 3D-RANSAC maps at FSC = 0.5 and FSC = 0.143 are 19 and 12 Å, respectively, which means that both structures are similar, as visually suggested already in Figure 6c, certainly enough for any of them to be used as a low-resolution initial 3D map. However, the processing time for obtaining the 3D-RANSAC map with the same parameters and the same laptop computer than in previous cases is only 50 min. The processing time of CL2D is of 435 min, and then, the total time

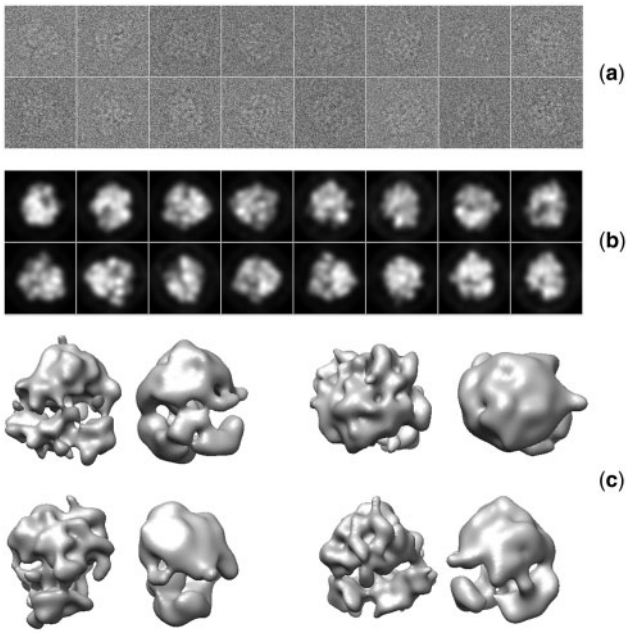


Fig. 6. Eight experimental projections of the eukaryotic ribosome (**a**), 16 class average images low-pass filtered (**b**) and 3D map obtained by PRIME approach (left) together with our best obtained initial volume (right) (with highest score) at four different orientations (**c**)

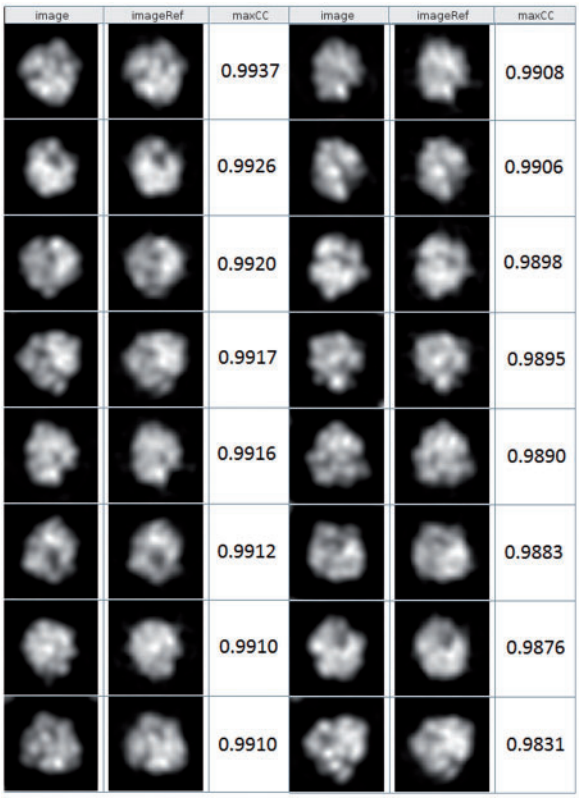


Fig. 7. Experimental class averages (image) of the asymmetric eukaryotic ribosome particles, corresponding initial volume projection at the same orientations (imageRef) and normalized cross correlation between both images (maxCC)

CL2D + 3D-RANSAC is of 485 min, whereas PRIME (or virtually any other method so far proposed) requires more than 4 days of computation (5760 min) in a laptop setting. To present the good agreement between the obtained initial volume and the used class averages, we show in Figure 7, the experimental class averages (labeled as 'image'), the corresponding initial 3D map projection at the same orientation ('imageRef') and the normalized cross correlation between both images ('maxCC'). As can be seen from Figure 7, there is a good agreement between the class averages and the corresponding projections.

3.2.3 Case 3: GroEL Additionally, we have used a GroEL dataset, kindly made available by Dr Ludtke (Ludtke *et al.*, 2004), composed by 26 micrographs of size 4082×6278 pixels. The sampling rate is 2.10 Å/pixel and the microscope voltage is of 200 kV. From this dataset, we have detected 4123 particles of size 128×128, using the method presented in (Abrishami *et al.*, 2013), and 16 classes were determined using CL2D (Sorzano *et al.*, 2010). In Figure 8a, we show the obtained 16 classes and the best volume (b) recovered by 3D-RANSAC map using default parameters, as before. In this case, the required processing time is ~5 min with the same computer as before. The processing time of CL2D is of 318 min, and then, the total time CL2D + 3D-RANSAC is of 324 min. The FSC curves between the phantom EMDB (EMDB code 1081, Ludtke *et al.*, 2004) and the 3D-RANSAC map at FSC = 0.5 and FSC = 0.143 are 10.2 and 9.7 Å, respectively, which means that both structures are similar. To further show the good agreement between the obtained initial volume and the used class averages, we present in Figure 9, the experimental class averages in the first row labeled as 'image', the corresponding initial volume projections at the same orientation in the second row ('imageRef') and the normalized cross correlation between both images in the third row ('maxCC'). As can be seen from Figure 9, there is a good agreement between the class averages and the corresponding projections.

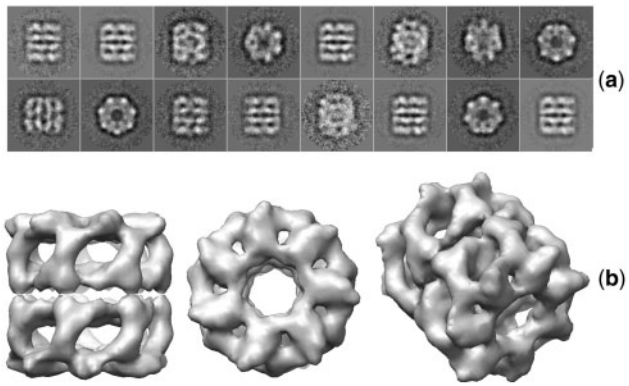


Fig. 8. Sixteen experimental class average images obtained using CL2D of GroEL experimental projections (**a**) and the best volume recovered by the proposed method at three different orientations (**b**)

3.2.4 Case 4: MCM467 Finally, for the sake of completeness, we show the behavior when dealing with small complexes in the

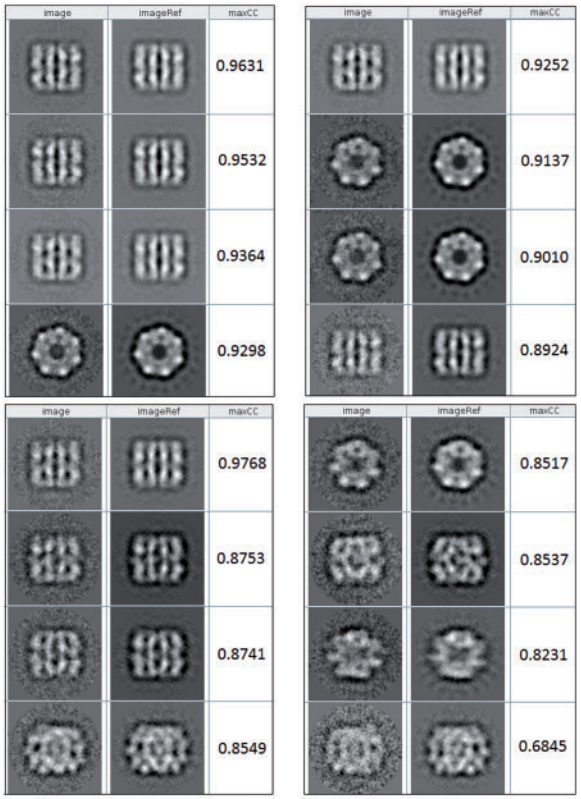


Fig. 9. Experimental class average images of the GroEL particles (image), corresponding initial volume projection at the same orientations (imageRef) and normalized cross correlation between both images (maxCC)

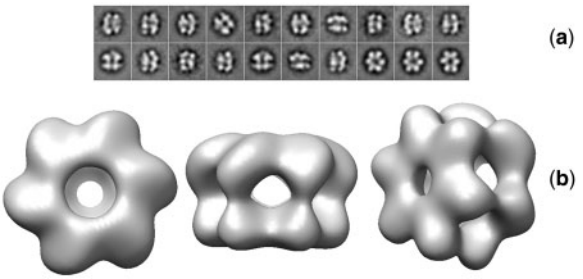


Fig. 10. Twenty obtained classes from experimental projections of MCM467 complex using CL2D approach (**a**) and three different orientations of the best initial volume obtained by the proposed method (**b**)

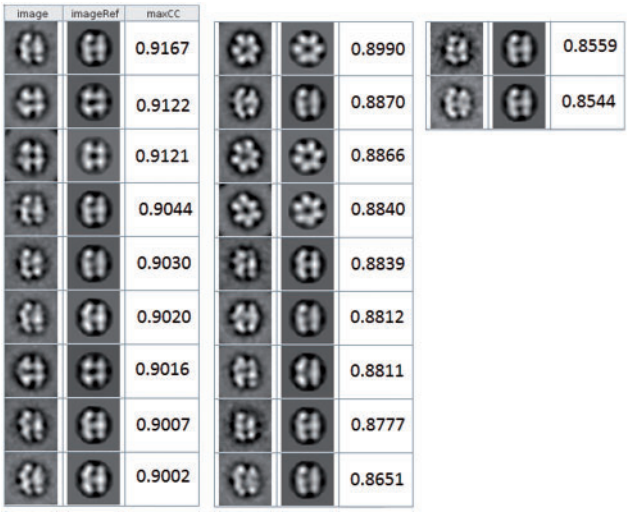


Fig. 11. Experimental class averages of the MCM467 complex (image), corresponding initial volume projection at the same orientations (imageRef) and normalized cross correlation between both images (maxCC)

order of a half million daltons using negative staining. This is an internally acquired dataset, corresponding to a certain MCM467 complex. The microscope is a JEOL JEM-1230, and the accelerating voltage is 100 kV. The nominal magnification is ×40 000 and the sampling rate 2.28 Å/pixel. We have obtained ∼6000 particles from 200 micrographs, using the method presented in (Abrishami *et al.*, 2013). From the picked particles, we obtained 128 class averages using CL2D (Sorzano *et al.*, 2010), which are manually curated to a smaller homogeneous dataset of 20 class averages.

In Figure 10a, we show the obtained class averages, after the curation process. As in the cases shown before, we have used the same input parameters for the determination of the 3D-RANSAC map. In this case, the required processing time is of 6 min with the same computer as before. The processing time of CL2D is of 720 min, and then, the total time CL2D + 3D-RANSAC is of 726 min. In Figure 10b we show the 3D-RANSAC map at different orientations.

As in the case before, to show the good agreement between the obtained initial volume and the used class averages, we present in Figure 11, the class averages, the corresponding projection of the initial volume at the same orientation and the normalized cross correlation between both images. As can be seen from Figure 11, there is a good agreement between the class averages and the corresponding projections.

4 DISCUSION AND CONCLUSION

Obtaining a reliable low-resolution initial map is a well-known current challenging problem in single particle electron microscopy. This statement can easily be supported considering the large number of recent publications about this topic (Elmlund and Elmlund, 2012; Elmlund *et al.*, 2013; Lyumkis *et al.*, 2013; Voss *et al.*, 2010). Existing current approaches are mainly based on common lines (Castón *et al.*, 1999; Crowther *et al.*, 1970; Elmlund and Elmlund, 2012; Elmlund *et al.*, 2010; Thuman-Commike and Chiu, 1997) or random model generation (Harauz and van Heel, 1985; Sanz-Garcia *et al.*, 2010; van Heel, 1984). However, in general, these methods are not easy to use, in term of the selection of input parameters, and are computing intensive. Naturally, the combination of a non-easy choice of input parameters together with long execution times complicates their practical use, requiring expert processing. Aware of these problems, we have set our aim at developing a simple-to-use method for which default parameters work well in most cases, at the same time as when the required computing time is minimized to minutes on a standard laptop

computer. Naturally, most methods can be trapped into local minima, and 3D-RANSAC is not an exception, but in this case, the landscape of solutions is particularly smooth, the use of RANSAC algorithm and, additionally, its short execution time reduce considerably this risk and also open the venue to future developments involving close-to-global optimization techniques for particularly challenging problems. Finally, we have developed 3D-RANSAC for the case of homogenous image populations, and at this stage this may be considered a limitation of the method. Observe that the input of the algorithm are homogeneous class average images, and therefore, the problem of separating structurally heterogeneous image sets into homogeneous classes has to be already solved in the previous classification approach. The extension of 3D-RANSAC to the non-homogenous case will be considered in future works, requiring a far from trivial extension of many concepts behind RANSAC.

In this article, we have presented a fast and efficient approach to obtain a reliable low-resolution initial volume from sets of macromolecular projection images without a priori information. The proposed method, instead of trying to explore the entire space of projection orientations, a task that is computationally intractable even for a few hundred images, uses a novel random modeling strategy based on an initial non-linear dimensionality reduction and RANSAC algorithm, which makes this approach efficient from a computational point of view. Observe that the probability of assigning the correct orientation to one projection corresponds to a standard uniform distribution. As the process of assigning the orientation to projections is statistically independent, the probability of giving correctly the angles to n particles corresponds to p^n (with P the probability to assign correctly the orientation to one projection). Therefore, if the number of images is high, this probability is low. Therefore, to increase this probability, we have used smaller sets of images (of typically nine projections) but at the same time capturing most of the structural information of the volume (note that this process is accomplished by the dimensionality reduction approach). 3D-RANSAC is a two-step survival algorithm. In the first step, a large number of models are generated (~380) but only the best k survive, which are the ones with the highest number of inliers (largest score). After this first selection process, these k models are refined again using all the initial classes by a projection matching approach, and at the end of this second selection step they are ranked again so that there is only one winner. We have tested our proposed method with synthetic (Bacteriorhodopsin) and experimental data (*BPV*, Eukaryotic Ribosome, GroEL and MCM467). In all cases, we have obtained fast and satisfactory results. The algorithm is freely available as a part of the Xmipp 3.1 package (de la Rosa-Trevín *et al.*, 2013).

Funding: The authors would like to acknowledge economical support from the Spanish Ministry of Economy and Competitiveness through grants AIC–A–2011–0638 and BIO2010-16566, the Comunidad de Madrid through grant CAM(S2010/BMD-2305) and the NSF through grant 1114901, as well as postdoctoral 'Juan de la Cierva' grant with reference JCI-2011-10185. C.O.S. Sorzano is recipient of a Ramón y Cajal fellow.

Conflict of interest: none declared.

REFERENCES

Abrishami,V. *et al.* (2013) A pattern matching approach to selection of particles from low-contrast electron micrographs. *Bioinformatics*, **29**, 2460–2468.

Bilbao-Castro,J.R. *et al.* (2004) Phan3D: design of biological phantoms in 3D electron microscopy. *Bioinformatics*, **20**, 3286–3288.

Baker,T.S. and Cheng,R.H. (1996) A model-based approach for determining orientations of biological macromolecules imaged by cryo-electron microscopy. *J. Struct. Biol.*, **116**, 120–130.

Cantele,F. *et al.* (2003) The variance of icosahedral virus models is a key indicator in the structure determination: a model-free reconstruction of viruses, suitable for refractory particles. *J. Struct. Biol.*, **141**, 84–92.

Castón,J.R. *et al.* (1999) A strategy for determining the orientations of refractory particles for reconstruction from cryo-electron micrographs with particular reference to round, smooth-surfaced, icosahedral viruses. *J. Struct. Biol.*, **125**, 209–215.

Crowther,R.A. *et al.* (1970) Three dimensional reconstructions of spherical viruses by fourier synthesis from electron micrographs. *Nature*, **226**, 421–425.

de la Rosa-Trevín,J.M. *et al.* (2013) Xmipp 3.0: An improved software suite for image processing in electron microscopy. *J. Struct. Biol.*, **13**, 256–256.

Elmlund,D. *et al.* (2010) Ab initio structure determination from electron microscopic images of single molecules coexisting in different functional states. *Structure*, **18**, 777–786.

Elmlund,H. *et al.* (2008) A new cryo-EM single-particle ab initio reconstruction method visualizes secondary structure elements in an ATP-fuelled AAA+ motor. *J. Mol. Biol.*, **375**, 934–947.

Elmlund,D. and Elmlund,H. (2012) SIMPLE: software for ab initio reconstruction of heterogeneous single-particles. *J. Struct. Biol.*, **180**, 420–427.

Elmlund,H. *et al.* (2013) PRIME: probabilistic initial 3D model generation for single-particle cryo-electron microscopy. *Structure*, **21**, 1299–1306.

Fischler,M.A. and Bolles,R.C. (1981) Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM*, **24**, 381–395.

Frank,J. (1996) *Three-Dimensional Electron Microscopy of Macromolecular Assemblies*. Academic, New York, NY.

Giannakis,D. *et al.* (2012) The symmetries of image formation by scattering. I. Theoretical framework. *Opt. Express*, **20**, 12799–12826.

Harauz,G. and van Heel,M. (1985) Direct 3D reconstruction from projections with initially unknown angles. In: Gelsema,E.S. and Kanal,L.N. (eds) *Pattern Recognition in Practice II*. Elsevier, North-Holland Publishing, Amsterdam, pp. 279–288.

Henderson,R. *et al.* (1990) Model for the structure of bacteriorhodopsin based on high-resolution electron cryo-microscopy. *J. Mol. Biol.*, **213**, 899–929.

Leschziner,A.E. and Nogales,E. (2006) The orthogonal tilt reconstruction method: an approach to generating single-class volumes with no missing cone for ab initio reconstruction of asymmetric particles. *J. Struct. Biol.*, **153**, 284–299.

Ludtke,S.J. *et al.* (2004) Seeing GroEL at 6 Å resolution by single particle electron cryomicroscopy. *Structure*, **12**, 1–20.

Liu,X. *et al.* (2007) Averaging tens to hundreds of icosahedral particle images to resolve protein secondary structure elements using a multi-path simulated annealing optimization algorithm. *J. Struct. Biol.*, **160**, 11–27.

Lyumkis,D. *et al.* (2013) Optimod–an automated approach for constructing and optimizing initial models for single-particle electron microscopy. *J. Struct. Biol.*, **184**, 417–426.

Ogura,T. and Sato,C. (2006) A fully automatic 3D reconstruction method using simulated annealing enables accurate posterioric angular assignment of protein projections. *J. Struct. Biol.*, **156**, 371–386.

Penczek,P.A. *et al.* (1996) A common-lines based method for determining orientations for N > 3 particle projections simultaneously. *Ultramicroscopy*, **63**, 205–218.

Radermacher,M. *et al.* (1987) Three-dimensional reconstruction from a single-exposure random conical tilt series applied to the 50S ribosomal subunit of *Escherichia coli*. *J. Microsc.*, **146**, 113–136.

Sanz-Garcia,E. *et al.* (2010) The random-model method enables ab initio 3D reconstruction of asymmetric particles and determination of particle symmetry. *J. Struct. Biol.*, **171**, 216–222.

Scheres,S.H.W. *et al.* (2007) Disentangling conformational states of macromolecules in 3D-EM through likelihood optimization. *Nat. Methods*, **4**, 27–29.

Sorzano,C.O.S. *et al.* (2010) A clustering approach to multireference alignment of single-particle projections in electron microscopy. *J. Struct. Biol.*, **171**, 197–206.

Thuman-Commike,P.A. and Chiu,W. (1997) Improved common-line-based icosahedral particle image orientation estimation algorithms *Ultramicroscopy*, **68**, 231–255.

Taylor,K.A. and Glaeser,R.M. (2008) Retrospective on the early development of cryo-electron microscopy of macromolecules and a prospective on opportunities for the future. *J. Struct. Biol.*, **163**, 214–223.

Voss,N.R. *et al.* (2010) A toolbox for ab initio 3-D reconstructions in single-particle electron microscopy. *J. Struct. Biol.*, **169**, 389–398.

Wolf,M. *et al.* (2010) Subunit interactions in bovine papillomavirus. *Proc. Natl Acad. Sci. USA*, **107**, 6298–6303.

van Heel,M. (1987) Angular reconstitutiona posteriori assignment of projection directions for 3-D reconstruction. *Ultramicroscopy*, **21**, 111–123.

van Heel,M. (1984) Three-dimensional reconstruction from projections with unknown angular relationships. In: Csanády,Á. *et al.* (ed.) *Eighth European Congress on Electron Microscopy*. Vol. 2. Programme Committee of the Eighth European Congress on Electron Microscopy, Budapest, Hungary, pp. 1347–1348.

Zhang,Z. and Hongyuan,Z. (2005) Principal manifolds and nonlinear dimension reduction via local tangent space alignment. *SIAM J. Sci. Comput.*, **26**, 313–338.

Zhang,X. and Zhou,H.Z. (2011) Limiting factors in atomic resolution cryo electron microscopy: no simple tricks. *J. Struct. Biol.*, **175**, 253–263.

Journal of Structural Biology 189 (2015) 213–219

Contents lists available at ScienceDirect

Journal of Structural Biology

journal homepage: www.elsevier.com/locate/yjsbi

A statistical approach to the initial volume problem in Single Particle Analysis by Electron Microscopy

C.O.S. Sorzano [a,b,*], J. Vargas [a], J.M. de la Rosa-Trevín [a], J. Otón [a], A.L. Álvarez-Cabrera [a], V. Abrishami [a], E. Sesmero [b], R. Marabini [c], J.M. Carazo [a]

[a] National Center of Biotechnology (CSIC), c/Darwin, 3, Campus Univ. Autonoma de Madrid, 28049 Cantoblanco, Madrid, Spain
[b] Bioengineering Lab., Univ. San Pablo CEU, Campus Urb. Monteprincipe s/n, 28668 Boadilla del Monte, Madrid, Spain
[c] Escuela Politecnica Superior, Univ. Autonoma de Madrid, Campus. Univ. Autonoma de Madrid, 28049 Cantoblanco, Madrid, Spain

ARTICLE INFO

Article history:
Received 22 October 2014
Received in revised form 30 December 2014
Accepted 17 January 2015
Available online 28 January 2015

Keywords:
3D reconstruction
Initial volume

ABSTRACT

Cryo Electron Microscopy is a powerful Structural Biology technique, allowing the elucidation of the three-dimensional structure of biological macromolecules. In particular, the structural study of purified macromolecules –often referred as Single Particle Analysis(SPA)– is normally performed through an iterative process that needs a first estimation of the three-dimensional structure that is progressively refined using experimental data. It is well-known the local optimisation nature of this refinement, so that the initial choice of this first structure may substantially change the final result. Computational algorithms aiming to providing this first structure already exist. However, the question is far from settled and more robust algorithms are still needed so that the refinement process can be performed with sufficient guarantees.

In this article we present a new algorithm that addresses the initial volume problem in SPA by setting it in a Weighted Least Squares framework and calculating the weights through a statistical approach based on the cumulative density function of different image similarity measures. We show that the new algorithm is significantly more robust than other state-of-the-art algorithms currently in use in the field.

The algorithm is available as part of the software suite Xmipp (http://xmipp.cnb.csic.es) and Scipion (http://scipion.cnb.csic.es) under the name "Significant".

1. Introduction

Single Particle Analysis using the Electron Microscope is a powerful experimental technique to elucidate the three-dimensional structure of macromolecular complexes (Frank, 2006; Sorzano et al., 2007). Thousands of two-dimensional projections of the structure under study are collected with the Electron Microscope, which are then used in most cases within iterative algorithms that have as initial input a first estimation of the three-dimensional structure. However, refinement algorithms are known to behave as local optimizers (Sorzano et al., 2006; Henderson et al., 2012), so that the dependence of the final result on the initial volume is a major concern in the field. This situation is known as the "initial volume problem". There exist several algorithms addressing the task of reconstructing a 3D volume compatible either with the 2D experimental images or with their image class averages (Penczek et al., 1996; Ogura and Sato, 2006; Singer et al., 2010; Coifman et al., 2010; Elmlund et al., 2010; Sanz-García et al., 2010; Singer and Shkolnisky, 2011; Elmlund and Elmlund, 2012; Elmlund et al., 2013; Vargas et al., 2014). However, the problem is far from settled due to several reasons: (1) It is an optimisation problem in a high-dimensional space; (2) There are many local minima and algorithms may get trapped into them. Except for Elmlund et al. (2013), most algorithms aim at trying to avoid local minima. Elmlund et al. (2013) takes a soft optimisation probabilistic approach, in which an image can take multiple 3D orientations with different weights calculated from some heuristically determined function within a subset of so-called feasible directions. This idea is somehow similar to the one in Maximum Likelihood and Bayesian reconstruction (Scheres et al., 2005, 2007; Scheres, 2012a), in which all projections can take all directions with different weights (in this case, calculated from the assumed *a priori* distribution of noise (ML) and signal coefficients (Bayesian)). In turn, Vargas et al. (2014) adopts a statistical approach with the goal of

* Corresponding author at: National Center of Biotechnology (CSIC), c/Darwin, 3, Campus Univ. Autonoma de Madrid, 28049 Cantoblanco, Madrid, Spain. Fax: +34 91 585 4506.
E-mail address: coss@cnb.csic.es (C.O.S. Sorzano).

http://dx.doi.org/10.1016/j.jsb.2015.01.009
1047-8477/

also avoiding the local minima by strongly reducing the search space using image subsets, randomly assigning Euler angles and checking which of the assignments was more successful. Unfortunately, current practice shows that, despite the availability of all these possibilities, more robust algorithms are still in need, since there are occasions in which the existing programs fail to produce a satisfactory result. Some recent approaches (like Optimod (Lyumkis et al., 2013) or MyFirstMap) take the pragmatic approach of generating many different volumes (preferably with different algorithms) and rank the volumes according to their fit to the experimental data.

The algorithm presented in this paper, which we will refer to as Significant, follows previous approaches in the field in which an image is allowed to have different projection directions with different weights. However, instead of setting the problem as a closed form optimisation of a given functional under a simplified set of assumptions, which may be violated in practical works, it considers more realistic models at the expense of mathematical tractability. We rely on the theory of Weighted Least Squares (WLS) optimisation rather than, for instance, on Maximum Likelihood (ML) optimisation. The rationale for this choice is that we are more free to choose a different weight scheme in which we incorporate more criteria evaluating the quality of the fitting between a given particle and its candidate projection direction. The fact that the functional is changed along iterations complicates its mathematical properties in the limit, so that the algorithm cannot be understood as an iterative algorithm to solve a Weighted Least Squares problem because the weights change from iteration to iteration. In principle, no weighting scheme is better than another, and the proof of its correctness can only be based on the results it produces.

Following the rational just introduced, Significant has been developed so that similarity measures are certainly addressed within statistically significant intervals; additionally, we have incorporated a number of new "desired properties" of a solution. In this way, we introduce the notion of "images being important for a projection" and of "projections being important for an image", the explicit consideration of the spatial neighbourhood of projection directions and, finally, the combined use of several image similarity measures (the correlation coefficient and the IMED (IMage Euclidean Distance) (Wang et al., 2005), an image metric that takes into account pixel neighbourhoods). In the Results section we compare our new algorithm with a number of common methods in the field.

2. Methods

Let us call I_i the ith image in a collection of N images (they can be experimental images or class averages, from the point of view of our algorithm the only difference is a larger execution time in the case of experimental images, since there are many more experimental images than class averages). In order to construct a first reference volume, we assign random angles to each one of the images and make a first reconstruction, that we will refer to as $V^{(0)}$. This first reconstruction normally looks as a smooth sphere whose radius coincides with the particle radius. If a better prior exists (the volume is approximately a cylinder, or even a previous 3D reconstruction of a related molecule), we may use it instead.

Let us now refine the first reconstruction using the following iterative method

$$V^{(k+1)} = \arg\min_V \sum_{i=1}^{N}\sum_{j=1}^{M} w_{ij}^{(k)} \|\widetilde{I}_{ij}^{(k)} - P_j V\|^2 \quad (1)$$

where P_j denotes the projection operator along the direction j (assuming that we are exploring a discrete library of M projections), and $\widetilde{I}_{ij}^{(k)}$ is the image resulting of aligning, rotationally and translationally, the ith image to the jth projection of $V^{(k)}$. $w_{ij}^{(k)}$ is a weight (note that normally weights are between 0 and 1, and this is indeed the case in our method, although this is not strictly necessary) that controls whether the ith image should be considered to come from the jth direction at iteration k. Note that many of the 3D reconstruction formalisms can be set in this generic framework: Projection Matching (Scheres et al., 2008) has $w_{ij}^{(k)} = 1$ for only one of the M directions; in Maximum-Likelihood 3D (Scheres et al., 2007) all weights can, in principle, be different from 0 and they are calculated based on the *a priori* assumption of Gaussianly distributed noise; similarly, Relion calculates weights based on the previous assumption and the assumption that Fourier coefficients are Gaussianly distributed (Scheres, 2012a). This type of algorithms is referred as Weighted Least Squares (WLS).

In this article, we also adopt a probabilistic approach for the weight calculation, although in this case based on the concept of statistical significance. Let us consider the case of Projection Matching. It compares, after alignment, the ith image to all M projections generated from the volume at iteration k. This comparison is usually performed by calculating Pearson's correlation coefficient between the two images, $\rho_{ij}^{(k)}$, and the algorithm selects the direction with maximum correlation. However, since images are noisy, the correlation coefficient itself is a random variable. If both the experimental images and the reprojections were to follow a normal distribution, the one-sided confidence interval associated to their cross correlation could be easily computed through Fisher's transformation (Sheskin, 2004, Chap. 28)

$$\rho \in \left[\tanh\left(\tanh^{-1}\left(\max_j\{\rho_{ij}^{(k)}\}\right) - \frac{z_{1-\alpha^{(k)}}}{\sqrt{N-3}}\right), \max_j\{\rho_{ij}^{(k)}\}\right] \quad (2)$$

where tanh is the hyperbolic tangent, α is the level of confidence, $z_{1-\alpha^{(k)}}$ is the $1-\alpha^{(k)}$ percentile of the Gaussian distribution, and N is the number of pixels on which the correlation has been calculated. The idea is that, because of the noise, all those directions whose correlation coefficient lay in this confidence interval are statistically indistinguishable from the maximum (with a confidence level $\alpha^{(k)}$), and consequently, they should all be kept as feasible solutions. However, the assumption of normality does not hold in practical cases (this issue will be further discussed along this work), which makes inaccurate the simple computation of Fisher's transformation. At this point Significant departs from other algorithms in the field in that it still uses Fisher's confidence interval as a first way to filter out direction candidates, but it subsequently explicitly considers the distribution of experimental correlation coefficients for the actual confidence assignment (note that this approach allows the use of other similarity measures besides cross correlation). This latter concept is what we will refer as "a direction being significant to an image" (with a confidence level $\alpha^{(k)}$). For doing so, we estimate the marginal probability density function of the $\rho_{i\cdot}^{(k)}$ variable (see Fig. 1), and we check whether $\rho_{ij}^{(k)}$ is larger than the $1-\alpha^{(k)}$ percentile:

$$\Pr\left\{\rho_{i\cdot}^{(k)} \leqslant \rho_{ij}^{(k)}\right\} \geqslant 1-\alpha^{(k)} \quad (3)$$

Note that in this condition $\alpha^{(k)}$ plays a similar role to the Type I error (α) in Statistical Inference, and from that analogy we have chosen the name "Significant" for this method. Note that the role of this condition is to allow the contribution of an image to a number of "Significant" directions at the same time, while working with the experimental distribution of similarity measures, without being restricted to normality assumptions or the use of cross correlations.

We may also add the desired condition that the image is significant to the direction by testing whether

$$\Pr\left\{\rho_{\cdot j}^{(k)} \leqslant \rho_{ij}^{(k)}\right\} \geqslant 1-\alpha^{(k)} \quad (4)$$

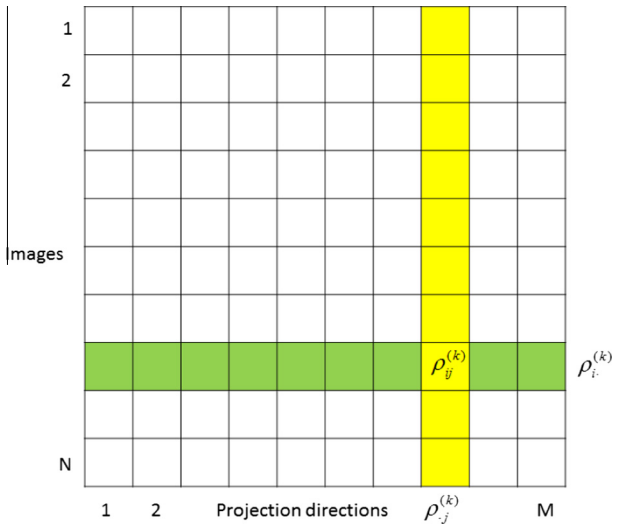


Fig.1. Graphical representation of the ρ_{ij} matrix and the marginal variables $\rho_{i\cdot}$ (the set of correlations for a given image) and $\rho_{\cdot j}$ (the set of correlations for a given direction).

(see Fig. 1). This condition may be used if we expect to have many outliers (empty images, images with no specimen,...), although its effects will naturally imply the selection of only a few images per projection direction and the rejection of the rest. We refer to this condition as the "strict" direction condition (the direction becomes "strict" about which images can contribute to it). In practical cases the "strict" conditions will be seldom used, with the exception of heavily contaminated data sets, for which "strict" can be very useful, as we will show in subsequent sections.

If the condition on direction significance is met, for which we will demand both the fulfilment of Fisher's condition and of the experimentally-determined significance interval, the weight is calculated as

$$w_{ij}^{(k)} = \frac{\rho_{ij}^{(k)}}{\max\limits_{\substack{i'\in 1,\dots,N\\ j'\in \text{Neigh}_\theta(j)}} \rho_{i'j'}^{(k)}} \Pr\left\{\rho_{i\cdot}^{(k)} \leqslant \rho_{ij}^{(k)}\right\}\Pr\left\{\rho_{\cdot j}^{(k)} \leqslant \rho_{ij}^{(k)}\right\} \tag{5}$$

where $\text{Neigh}_\theta(j)$ is the set of projection directions that differ from the jth projection direction in less than θ degrees; otherwise, the weight is zero. Note that the use of $\text{Neigh}_\theta(j)$ represents a new "criterion" of our solution: its correlation should be a good match also when compared to its surroundings. The rationale is that if an image is correctly assigned to a given direction, then its correlation should also be amongst the best in a neighbourhood of that direction. Note that Projection Matching is obtained in this scheme if $\alpha = \frac{1}{N}$ and we drop any reference to the distribution of correlations along a direction ($\rho_{\cdot j}$). In the Supplementary Material we show the effect of the conditions used in this method to select candidate directions.

Of course, if the condition of "strict" direction is enabled, we will also reinforce that images with similarity measures significantly lower than the maximum value will be assigned a weight equal to zero.

Cross-correlation among two different images is simply a way to measure their similarity. There have been other proposals to measure this similarity like IMED (Wang et al., 2005). IMED is, in fact, a generalisation of the Euclidean distance between two images X and Y that takes into account local neighbourhoods of each pixel:

$$\eta(X,Y) = \sum_{m=1}^{P}\sum_{n=1}^{P} \exp\left(\frac{\|\mathbf{r}_m - \mathbf{r}_n\|^2}{2}\right)(X(\mathbf{r}_m) - Y(\mathbf{r}_m))(X(\mathbf{r}_n) - Y(\mathbf{r}_n)) \tag{6}$$

where P is the number of pixels in images X and Y, $\mathbf{r}_n$ denotes the location of the nth pixel, and $X(\mathbf{r}_n)$ the pixel value at that location. We have observed that IMED has a better discriminatory power than cross-correlation. For instance, Fig. 2 shows a plot of the cross-correlation and IMED at the high-end of cross-correlation (images that are very similar to each other according to cross-correlation). The line plotted is a polynomial of degree 3 fitted to this data. Note the increase of slope of IMED with respect to cross-correlation at very high correlation values revealing the more discriminatory power of IMED (note also that IMED values decrease as cross-correlation values increase).

Finally, we calculate the weight as

$$w_{ij}^{(k)} = \left(\frac{\min\limits_{\substack{i'\in 1,\dots,N\\ j'\in \text{Neigh}_\theta(j)}} \eta_{i'j'}^{(k)}}{\eta_{ij}^{(k)}} \Pr\left\{\eta_{i\cdot}^{(k)} \geqslant \eta_{ij}^{(k)}\right\}\Pr\left\{\eta_{\cdot j}^{(k)} \geqslant \eta_{ij}^{(k)}\right\}\right) \left(\frac{\rho_{ij}^{(k)}}{\max\limits_{\substack{i'\in 1,\dots,N\\ j'\in \text{Neigh}_\theta(j)}} \rho_{i'j'}^{(k)}} \Pr\left\{\rho_{i\cdot}^{(k)} \leqslant \rho_{ij}^{(k)}\right\}\Pr\left\{\rho_{\cdot j}^{(k)} \leqslant \rho_{ij}^{(k)}\right\}\right) \tag{7}$$

Note that these weights are necessarily between 0 and 1, and that if an image is the best one for a given direction and that direction is the best for that image, then $w_{ij} = 1$ as in Projection Matching.

A new volume is reconstructed (Eq. (1)) for iteration $k+1$ using the just calculated weights, and the process is iterated for a fixed number of times. At each iteration we normally increase our level of confidence, $1-\alpha^{(k)}$, by using a monotonically decreasing sequence of $\alpha^{(k)}$ values. However, any other strategy could be used. Typically, our confidence levels range between 85% and 99.99%. At low confidence levels, Fisher's confidence interval is relatively small because we do not need to be very confident about it, while the number of candidate directions amongst the top $1-\alpha^{(k)}$ percentile is relatively large. As the confidence level increases, Fisher's confidence interval increases, to account for the larger confidence needed, while the number of candidate directions in the top list decreases (because the $1-\alpha^{(k)}$ percentile increases).

We may think of the new reconstruction algorithm as being half way between Projection Matching (only one direction has a weight different from zero) and Maximum Likelihood/Maximum *a posteriori*

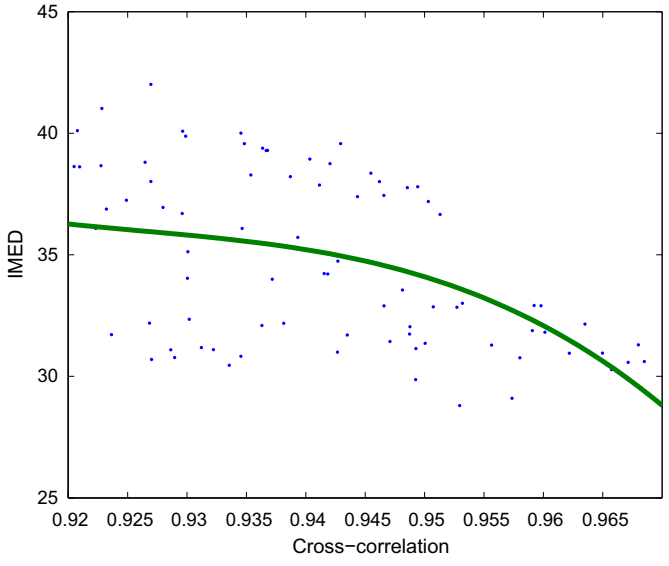


Fig.2. Scatter plot of cross-correlation and IMED values for the GroEL example. Only those images with cross-correlation higher than 0.92 are shown.

(all directions get a weight different from zero), but generalised to any type of similarity measures and to experimentally-determined similarity value distributions. In the new scheme, only a few directions get a non-zero weight (the number of projections range between 100 and 1, depending on the total number of projections, M, and the confidence level, $1-\alpha$), specifically those that are significantly more similar to each image. Interestingly, we have also introduced the notion of an image being significant for a direction. Typically, Electron Microscopy algorithms always assign a projection direction to any input image. However, in our scheme this may not be the case if the image at hand is not good enough (meeting our confidence conditions) for any of the projection directions. This prevents images with very low Signal-to-Noise Ratios, empty images and images corresponding to sufficiently different conformations from being used during the reconstruction.

3. Results

3.1. GroEL

GroEL (Ranson et al., 2001) is considered to be a difficult case for blind initial volume algorithms, because its top views and side views are approximately of the same size, and the algorithm does not always find their correct relative orientation. We used the GroEL dataset publicly available as the tutorial of EMAN2 (http://blake.bcm.edu/emanwiki/Ws2011/Eman2) (Tang et al., 2007). We automatically picked 8758 particles from 26 micrographs at a sampling rate of 4.2 Å/pixel using the algorithm described at Abrishami et al. (2013). We automatically evaluated their quality (Vargas et al., 2013) and kept 8589. Then, we performed a 2D classification (Sorzano et al., 2010) into 44 classes as a way to construct a "summary" of the collected data (see Fig. 3). It can be seen that some of the classes are of much better quality than others, with a quite different number of images assigned to them, as is normally the case in an experimental setting.

We compared the results of reconstruct_significant (run with 100 iterations and α linearly decreasing from 0.15 to 0.001; with a non-strict direction condition) to the results of EMAN2 (*e2initialmodel.py* run with 8 iterations), Simple 1.0 (Elmlund and Elmlund, 2012) (*origami* with low pass filtered to 15 Å, note that Simple 1.0 was originally introduced for raw images but that we are applying here to classes), RANSAC (Vargas et al., 2014) (run with 380 RANSAC iterations and an inlier threshold of 0.77), Relion (with autorefine) (Scheres, 2012b) and Projection Matching as implemented in Xmipp (Scheres et al., 2008). All these algorithms were run with their default parameters normally used in Xmipp. For those algorithms needing a starting volume, we constructed a "sphere" by assigning random angles to the 2D classes, performing a 3D reconstruction, and radially averaging the resulting volume. Relion cannot work with as few as 44 classes and we supplied it with the 8589 selected particles. Each algorithm produced 10 volumes (either by asking the algorithm to do so, or by repeating it 10 times) and found that the newly proposed algorithm constructed a correct model (understanding by correct a volume whose FSC = 0.5 frequency with respect to the EMDB volume is finer than 25 Å) 10 times, RANSAC 3 times, EMAN2 and Simple 1.0 2 times,

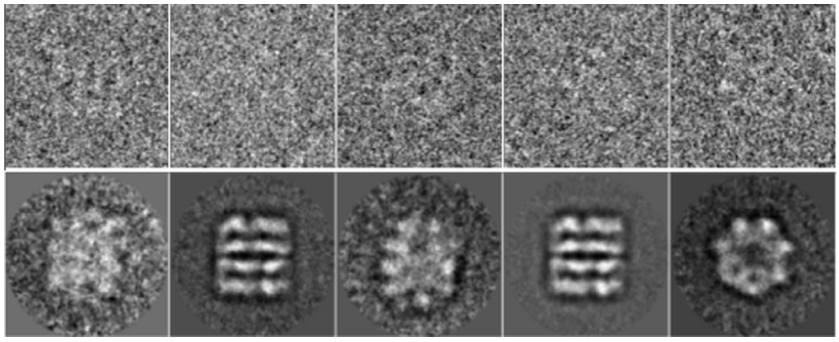


Fig.3. Sample images and classes of GroEL.

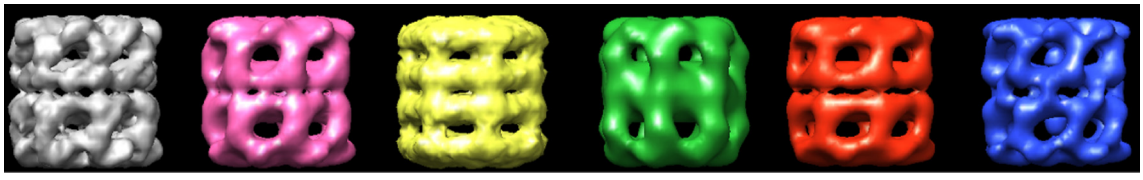


Fig.4. Examples of correct GroEL reconstructions. From left to right: GroEL structure deposited at EMDB (1042) at 10.3 Å; reconstruct_significant (cross-correlation, *cc*, to EMDB-1042: 0.841); Simple 1.0 (*cc* = 0.834); EMAN2 (*cc* = 0.825); Relion (*cc* = 0.809); RANSAC (*cc* = 0.786).

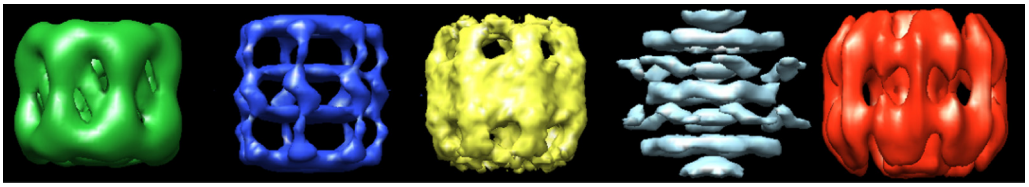


Fig.5. Examples of incorrect GroEL reconstructions. From left to right: EMAN2; RANSAC; Simple 1.0; Projection Matching; Relion.

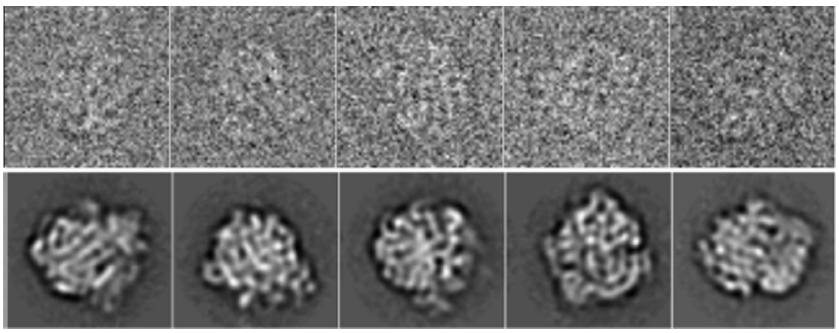


Fig.6. Sample images and classes of the eukaryotic ribosome.

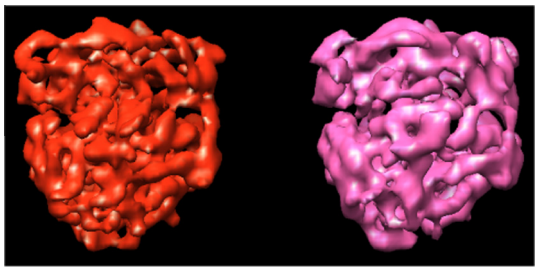


Fig.7. Examples of correct ribosome reconstructions. From left to right: Relion; reconstruct_significant (cross-correlation compared to Relion, $cc = 0.646$.).

Relion 1 time, Projection Matching 0 times. The execution time per volume using a single CPU was EMAN2 (2.7 min), RANSAC (2.8 min), Simple 1.0 (19 min), Projection Matching (25 min), reconstruct_significant (6 h), and Relion (33.6 days). Note that most of these algorithms are parallelized (including the newly proposed one) and the actual execution time must be divided by the number of processors available. Also, our algorithm produced a correct structure from iteration 4 ($\alpha = 0.856$) after only 24 min (in a single processor). Fig. 4 shows the correct GroEL structure as deposited at the Electron Microscopy Data Bank (entry 1042) and the best volumes reconstructed by each of the algorithms sorted by descending correlation coefficient while Fig. 5 shows some of the poorly reconstructed initial models.

In another experiment, we artificially added to the 44 class averages 396 images (=44 × 9) of pure noise with the same mean and standard deviation as the noise in the original images. Significant was capable of producing the correct structure if the "strict" direction condition was used (it was not without this condition). None of the other algorithms was able of producing a correct structure.

3.2. Eukaryotic ribosome

In our second experiment we have performed the blind *ab initio* reconstruction of 5000 cryo-EM projections of an eukaryotic ribosome, obtained from the EMDB test image data (http://www.ebi.ac.uk/pdbe/emdb/test_data.html) and originally used in the work of Scheres et al. (2007). The images had an original size of 130 × 130 pixels and we scaled them to a size of 64 × 64 pixels for speeding up the calculations. In our previous algorithm, RANSAC (Vargas et al., 2014), we needed to filter the 2D classes so that the algorithm was able to produce correct structures, the reason being probably that with high resolution information there were too many local minima in which the algorithm was getting trapped. In this experiment, we tested whether the new algorithm was able to produce good structures without any filtration and compared its results to the results of the rest of algorithms. Fig. 6 shows some of the images of the dataset and some of the classes calculated from them.

We estimated 32 2D classes using CL2D (Sorzano et al., 2010). We run the same set of programs as in the previous case, with the same parameters. This time, only Relion (only one run with the 5000 images) and Significant (in 100% of the 10 executions) were able to produce a correct structure (see Fig. 7). Fig. 8 shows the evolution along the iterations of the ribosome reconstructed by reconstruct_significant. The rest of the algorithms got trapped into local minima (see Fig. 9). The execution time per reconstructed model in a single CPU was: EMAN2 (3.4 min), Simple 1.0 (22 min), Projection Matching (36 min), RANSAC (10.5 h), Significant (16 h), Relion (37.3 days). Again, most of these algorithms are parallelized, so the actual wall clock time is much smaller.

To test whether the number of images played a role in this result, we provided EMAN2 and Simple 1.0 with the full set of raw images. None of the algorithms was capable of producing a good result after several days of execution.

4. Discussion

The determination of an initial volume that can be further refined in the context of iterative algorithms is a crucial step in the protocol of macromolecular structure determination from sets of Electron Micrographs. Practitioners in the field currently have a range of options, going from low-pass filtering a similar structure to *ab initio* 3D reconstruction passing by using random noise, geometrical models (Bilbao-Castro et al., 2004) and Random Conical

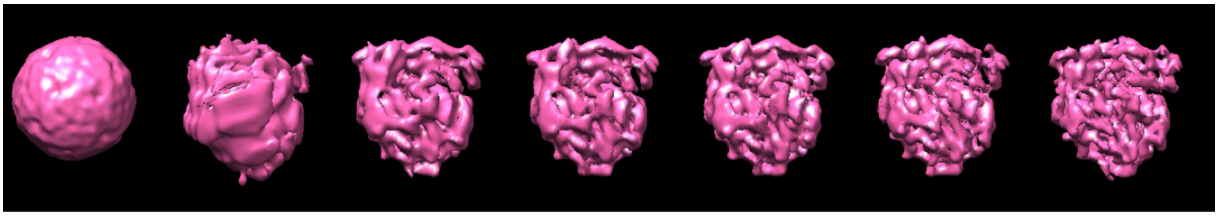


Fig.8. Evolution of the ribosome reconstruction along iterations (iteration 0, 15, 30, 45, 60, 75, and 90) using reconstruct_significant.

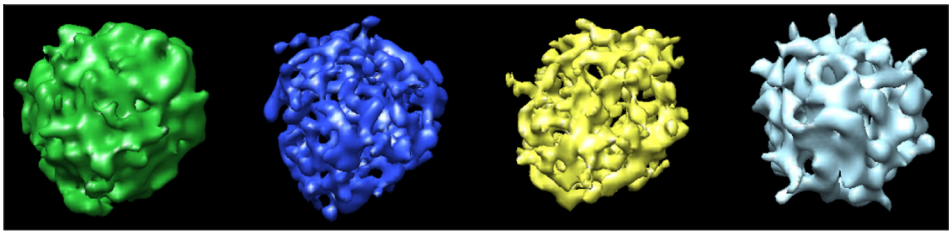


Fig.9. Examples of incorrect ribosome reconstructions. From left to right: EMAN2 ($cc = 0.297$); RANSAC ($cc = 0.217$); Simple 1.0 ($cc = 0.204$); Projection Matching ($cc = 0.196$).

Tilt reconstructions (Radermacher, 1988). *Ab initio* algorithms have extensively been explored by the EM community, as already shown in the introduction, with over 10 different methods. However, this large amount of possibilities may give the false impression that the problem is solved and that there is no need for yet another algorithm. However, reality is far from this. Indeed, there are many structures for which we can find an appropriate algorithm producing the correct result, however, there are also some other structures for which the existing algorithms still fall short. Even, as shown in our Results Section, there are structures for which a given algorithm may or may not produce a correct reconstruction. The algorithm proposed in this paper addresses these difficult situations. If the structure can be solved with fast methods like the one of EMAN2 or RANSAC, there is no need for an extra check with an algorithm such as the one proposed. However, if there are doubts about the plausibility of the initial model, running Significant may show worthy. Of course, we do not mean that the new algorithm is the definitive one, and there will always be room for improvement. However, in the Results Section we have shown that the algorithm is robust enough in situations with many (eukaryotic ribosome) or deep local minima (GroEL). The design of the algorithm borrows several ideas from many of the existing algorithms and combine them in a unique way with the aim of smoothing the landscape of solutions and finding the global minimum of the reconstruction problem. Specifically,

- It shares with Projection Matching and Maximum Likelihood (and partially with its Bayesian evolution in Relion) a Weighted Least Squares scheme. However, our weights are computed based on the cumulative distribution function of the correlation and IMED (a more robust distance measure between images) as well as the local structure of these two quantities around a given projection direction.
- It shares with Maximum Likelihood, Relion and Simple 1.0, the possibility that an image may contribute to multiple projection directions (with different weights). Again, there are differences with respect to the other algorithms in the way we choose those directions to which each image contributes to. We feel that our choice of comparing always each image to all possible projection directions (within the specified angular discretization limits) helps to not get trapped within local minima by taking decisions about the final projection direction too soon.
- It shares with Simple 1.0 and Simulated annealing the slow "cooling" scheme, in our case, the slow decrease of the Type I error (α). For each α we may think of the landscape of solutions of the modified Weighted Least Squares problem as one of a surrogate optimisation problem (we substitute the original landscape with many local minima by a smoother landscape with much fewer). As we get closer to the solution, we may go for fewer Type I errors (and consequently, more local minima). Obviously, reducing the number of iterations (in our examples, 100) for linearly going from α_0 to α_F would reduce the computing time, but it would also increase the risk of getting trapped into local minima. However, it is also true that the α sequence does not need to be linear and that faster sequences could be explored in the future as soon as we detect that we are in a sufficiently good minimum (which may occur rather early in the iterations, as was the case of GroEL).
- It introduces a new concept in EM that is the fact that the distribution of correlation and distance measures for a given projection direction also influences on the weight of the experimental image, and may even prevent it from participating in the 3D reconstruction ("strict" direction condition).

5. Conclusion

In this paper we have presented a new algorithm for the estimation of initial models that can be further refined by any of the already existing algorithms in EM. The algorithm is based on a Weighted Least Squares approach in which the weights are calculated using the cross correlation and IMED distance of the individual image to a specific projection direction as well as its relationship to neighbour directions and the comparison of these two quantities with respect to the rest of images available in the dataset (through the cumulative density functions defined in the Methods Section). All our design goes into the statistical direction of "being significant" (the projection direction must be significant for the image and vice versa, considering also the neighbourhood of that projection direction). We have experimentally shown that our algorithm succeeds in producing a correct initial guess in rather difficult cases. The conceptual bases of this new method can be expanded to other topics, such as 3D refinement at high resolution and 3D classification, although these extensions fall outside the scope of the present work.The algorithm is available through the open-source package Xmipp (http://xmipp.cnb.csic.es) (Sorzano et al., 2004; De la Rosa-Trevín et al., 2013) since version 3.2 (note that the official stable release is currently 3.1) under the name `xmipp_reconstruct_significant`.

Acknowledgments

Funding: The authors would like to acknowledge economical support from the Spanish Ministry of Economy and Competitiveness through Grants AIC-A-2011-0638, BFU2013-41249-P and BIO2013-44647-R, the Comunidad de Madrid through Grant CAM (S2010/BMD-2305), as well as postdoctoral Juan de la Cierva Grant with reference JCI-2011-10185. C.O.S. Sorzano is recipient of a Ramón y Cajal fellow.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.jsb.2015.01.009.

References

Abrishami, V., Zaldívar-Peraza, A., de la Rosa-Trevín, J.M., Vargas, J., Otón, J., Marabini, R., Shkolnisky, Y., Carazo, J.M., Sorzano, C.O.S., 2013. A pattern matching approach to the automatic selection of particles from low-contrast electron micrographs. Bioinformatics 29 (19), 2460–2468.

Bilbao-Castro, J.R., Sorzano, C.O.S., García, I., Fernández, J.J., 2004. Phan3D: design of biological phantoms in 3D electron microscopy. Bioinformatics 20, 3286–3288.

Coifman, R.R., Shkolnisky, Y., Sigworth, F.J., Singer, A., 2010. Reference free structure determination through eigenvectors of center of mass operators. Appl. Comput. Harmon. Anal. 28 (3), 296–312.

De la Rosa-Trevín, J.M., Otón, J., Marabini, R., Zaldívar-Peraza, A., Vargas, J., Carazo, J.M., Sorzano, C.O.S., 2013. Xmipp 30: one step forward in scientific computing for electron microscopy. J. Struct. Biol. 184, 321–328.

Elmlund, D., Elmlund, H., 2012. Simple: software for ab initio reconstruction of heterogeneous single-particles. J. Struct. Biol. 180 (3), 420–427.

Elmlund, D., Davis, R., Elmlund, H., 2010. Ab initio structure determination from electron microscopic images of single molecules coexisting in different functional states. Structure 18 (7), 777–786.

Elmlund, H., Elmlund, D., Bengio, S., 2013. Prime: probabilistic initial 3D model generation for single-particle cryo-electron microscopy. Structure 21 (8), 1299–1306.

Frank, J., 2006. Three-Dimensional Electron Microscopy of Macromolecular Assemblies: Visualization of Biological Molecules in Their Native State. Oxford Univ. Press, New York, USA.

Henderson, R., Sali, A., Baker, M.L., Carragher, B., Devkota, B., Downing, K.H., Egelman, E.H., Feng, Z., Frank, J., Grigorieff, N., Jiang, W., Ludtke, S.J., Medalia, O., Penczek, P.A., Rosenthal, P.B., Rossmann, M.G., Schmid, M.F., SchrÃüder, G.F., Steven, A.C., Stokes, D.L., Westbrook, J.D., Wriggers, W., Yang, H., Young, J., Berman, H.M., Chiu, W., Kleywegt, G.J., Lawson, C.L., 2012. Outcome of the first electron microscopy validation task force meeting. Structure 20 (2), 205–214.

Lyumkis, D., Brilot, A.F., Theobald, D.L., Grigorieff, N., 2013. Likelihood-based classification of cryo-em images using freealign. J. Struct. Biol. 183 (3), 377–388.

Ogura, T., Sato, C., 2006. Posterior Euler angle assignment using simulated annealing. J. Struct. Biol. 156, 371–386.

Penczek, P.A., Zhu, J., Frank, J., 1996. A common-lines based method for determining orientations for N > 3 particle projections simultaneously. Ultramicroscopy 63, 205–218.

Radermacher, M., 1988. Three-dimensional reconstruction of single particles from random and nonrandom tilt series. J. Electron Microsc. Tech. 9, 359–394.

Ranson, N., Farr, G., Roseman, A., Gowen, B., Fenton, W., Horwich, A., Saibil, H., 2001. ATP-bound states of GroEL captured by cryo-electron microscopy. Cell 107, 869–879.

Sanz-García, E., Stewart, A.B., Belnap, D.M., 2010. The random-model method enables ab initio 3D reconstruction of asymmetric particles and determination of particle symmetry. J. Struct. Biol. 171 (2), 216–222.

Scheres, S.H.W., 2012a. A Bayesian view on cryo-EM structure determination. J. Mol. Biol. 415 (2), 406–418.

Scheres, S.H.W., 2012b. Relion: implementation of a bayesian approach to cryo-em structure determination. J. Struct. Biol. 180 (3), 519–530.

Scheres, S.H.W., Valle, M., Núñez, R., Sorzano, C.O.S., Marabini, R., Herman, G.T., Carazo, J.M., 2005. Maximum-likelihood multi-reference refinement for electron microscopy images. J. Mol. Biol. 348, 139–149.

Scheres, S.H.W., Gao, H., Valle, M., Herman, G.T., Eggermont, P.P.B., Frank, J., Carazo, J.M., 2007. Disentangling conformational states of macromolecules in 3d-em through likelihood optimisation. Nat. Methods 4 (1), 27–29.

Scheres, S.H.W., Núñez Ramírez, R., Sorzano, C.O.S., Carazo, J.M., Marabini, R., 2008. Image processing for electron microscopy single-particle analysis using xmipp. Nat. Protocols 3, 977–990.

Sheskin, D.J., 2004. Handbook of Parametric and Nonparametric Statistical Procedures. Chapman & Hall/CRC.

Singer, A., Shkolnisky, Y., 2011. Three-dimensional structure determination from common lines in cryo-em by eigenvectors and semidefinite programming(). SIAM J. Imaging Sci. 4 (2), 543–572.

Singer, A., Coifman, R.R., Sigworth, F.J., Chester, D.W., Shkolnisky, Y., 2010. Detecting consistent common lines in cryo-em by voting. J. Struct. Biol. 169, 312–322 (under review).

Sorzano, C.O.S., Marabini, R., Velázquez-Muriel, J., Bilbao-Castro, J.R., Scheres, S.H.W., Carazo, J.M., Pascual-Montano, A., 2004. XMIPP: a new generation of an open-source image processing package for electron microscopy. J. Struct. Biol. 148, 194–204.

Sorzano, C.O.S., Marabini, R., Pascual-Montano, A., Scheres, S.H.W., Carazo, J.M., 2006. Optimization problems in electron microscopy of single particles. Ann. Oper. Res. 148, 133–165.

Sorzano, C.O.S., Jonic, S., Cottevieille, M., Larquet, E., Boisset, N., Marco, S., 2007. 3D electron microscopy of biological nanomachines: principles and applications. Eur. Biophys. J. 36, 995–1013.

Sorzano, C.O.S., Bilbao-Castro, J.R., Shkolnisky, Y., Alcorlo, M., Melero, R., Caffarena-Fernández, G., Li, M., Xu, G., Marabini, R., Carazo, J.M., 2010. A clustering approach to multireference alignment of single-particle projections in electron microscopy. J. Struct. Biol. 171, 197–206.

Tang, G., Peng, L., Baldwin, P.R., Mann, D.S., Jiang, W., Rees, I., Ludtke, S.J., 2007. Eman2: an extensible image processing suite for electron microscopy. J. Struct. Biol. 157, 38–46.

Vargas, J., Abrishami, V., Marabini, R., de la Rosa-Trevín, J.M., Zaldivar, A., Carazo, J.M., Sorzano, C.O.S., 2013. Particle quality assessment and sorting for automatic and semiautomatic particle-picking techniques. J. Struct. Biol. 183 (3), 342–353.

Vargas, J., Álvarez-Cabrera, A.L., Marabini, R., Carazo, J.M., Sorzano, C.O.S., 2014. Efficient initial volume determination from electron microscopy images of single particles. Bioinformatics 30, 2891–2898.

Wang, L., Zhang, Y., Feng, J., 2005. On the euclidean distance of images. IEEE Trans. Pattern Anal. Mach. Intell. 27, 1334–1339.

Structure

StructMap: Multivariate distance analysis of elastically aligned electron microscopy structures for exploring pathways of conformational changes

--Manuscript Draft--

Manuscript Number:	
Full Title:	StructMap: Multivariate distance analysis of elastically aligned electron microscopy structures for exploring pathways of conformational changes
Article Type:	Resource Article
Keywords:	Structure; dynamics; macromolecular complexes; electron microscopy; conformational changes; dissimilarity; distance; multivariate analysis; pathways
Corresponding Author:	Slavica Jonic CNRS-Université Pierre & Marie Curie Paris, FRANCE
First Author:	Carlos Oscar Sanchez Sorzano
Order of Authors:	Carlos Oscar Sanchez Sorzano
	Ana Lucia Alvarez-Cabrera
	Jose Maria Carazo
	Slavica Jonic
Abstract:	Single-particle electron microscopy has been shown to be very powerful for studying conformational flexibility of macromolecular complexes. To further analyze flexibility and start understanding dynamics of a complex, distinct conformations of this complex are usually computed by analyzing images of its coexisting multiple conformations. The resulting structures are then analyzed to explain flexibility. However, a quantitative analysis of dissimilarities (distances) among structures, placing the entire set of structures into a common space of comparison, is often lacking. Here, we present an approach that provides an overall view of distances among given structures. The approach is based on statistical analysis of distances among elastically aligned structures, and results in visualizing structures as points in a lower-dimensional distance space. The configuration of these points can be analyzed to explore potential pathways of conformational changes. The results of the method are shown with synthetic and experimental structures at different resolutions.
Suggested Reviewers:	Montserrat Samso Virginia Commonwealth University, Richmond, VA 23298, USA msamso@vcu.edu Expertise in EM dynamics studies of RyR1 (one of our experiments uses RyR1 structures deposited in EMDB by this expert).
	Oscar Llorca Centro de Investigaciones Biológicas - CSIC, Ramiro de Maeztu 9, 28040 Madrid, Spain ollorca@cib.csic.es Expertise in EM dynamics studies of Pol alpha –B subunit complex (one of our experiments uses structures of Pol alpha –B that were obtained in his team).
	Bridget Carragher The Scripps Research Institute, MC TRY-31, La Jolla, CA 92037, USA bcarr@scripps.edu Expertise in development and application of automated EM techniques for studying structure, function and dynamics of molecular machines.
	Florence Tama RIKEN, Kobe, Hyogo, 650-0047, Japan florence.tama@gmail.com Expertise in development and application of molecular mechanics simulation methods (the proposed methodology is based on normal mode analysis).

	Nikolaus Grigorieff Brandeis University, Waltham MA 02454-911 niko@grigorieff.org Expertise in development and application of EM methods for studying structure, function and dynamics of molecular machines.
	Andrej Sali University of California at San Francisco, San Francisco, CA 94158-2330, USA sali@salilab.org Expertise in structural bioinformatics methods and applications where the proposed methodology could also be useful.
Opposed Reviewers:	Abbas Ourmazd University of Wisconsin, Milwaukee, WI 53211, USA ourmazd@uwm.edu conflict of interest
	Joachim Frank HHMI, Columbia University, New York, NY 10032, USA jf2192@cumc.columbia.edu conflict of interest
	Sjors Scheres Medical Research Council, Cambridge Cb2 0qh, United Kingdom scheres@mrc-lmb.cam.ac.uk conflict of interest

Dr Slavica JONIC
IMPMC - UMR 7590
CNRS-University Pierre & Marie Curie
Campus Jussieu, Case courrier 115
4 Place Jussieu, 75252 Paris Cedex 05, France
Phone: +33 1 44 27 72 05
Fax: +33 1 44 27 37 85
E-mail: Slavica.Jonic@impmc.upmc.fr

To Editor of *Structure*

Paris, April 10, 2015

Dear Editor,

Please find attached the manuscript entitled "StructMap: Multivariate distance analysis of elastically aligned electron microscopy structures for exploring pathways of conformational changes" by C. O. S. Sorzano, A. L. Alvarez-Cabrera, J. M. Carazo, and S. Jonic, to be considered for publication as *Resource Article* in *Structure*.

Single-particle electron microscopy (EM) has been shown to be very powerful for studying conformational flexibility of large macromolecular complexes, by allowing computations of structures corresponding to coexisting different conformations of the same complex. However, a quantitative analysis of dissimilarities (distances) among structures, placing the entire set of structures into a common space of comparison, is often lacking. Here, we present an approach that provides an overall view of distances among given structures. The approach is based on statistical analysis of distances among elastically aligned structures, and results in visualizing structures as points in a lower-dimensional distance space. The configuration of these points can be analyzed to explore potential pathways of conformational changes, which is currently one of the major challenges in EM.

A list of recommended reviewers and the reviewers to exclude (conflict of interest) is given below. Thanks in advance for your consideration.

Yours sincerely,
Slavica Jonic

RECOMMENDED REVIEWERS:

1) **Montserrat Samso**
Department of Physiology and Biophysics
Virginia Commonwealth University, Richmond, VA 23298, USA
Email: **msamso@vcu.edu**
Reason for this suggestion: Expertise in EM dynamics studies of RyR1 (one of our experiments uses RyR1 structures deposited in EMDB by this expert).

2) **Oscar Llorca**
Centro de Investigaciones Biológicas - CSIC
Ramiro de Maeztu 9, 28040 Madrid, Spain
E-mail: **ollorca@cib.csic.es**
Reason for this suggestion: Expertise in EM dynamics studies of Pol alpha –B subunit complex (one of our experiments uses structures of Pol alpha –B that were obtained in his team).

3) **Bridget Carragher**
Department of Integrative Structural and Computational Biology
The Scripps Research Institute, MC TRY-31,
10550 North Torrey Pines Road, La Jolla, CA 92037, USA
Email: **bcarr@scripps.edu**
Reason for this suggestion: Expertise in development and application of automated EM techniques for studying structure, function and dynamics of molecular machines.

4) **Florence Tama**
RIKEN, Advanced Institute for Computational Sciences
7-1-26, Minatojima-minami-machi, Chuo-ku
Kobe, Hyogo, 650-0047, Japan
Email: **florence.tama@gmail.com**
Reason for this suggestion: Expertise in development and application of molecular mechanics simulation methods (the proposed methodology is based on normal mode analysis).

5) **Nikolaus Grigorieff**
MS 029, Rosenstiel Center
Brandeis University
415 South Street
Waltham MA 02454-911
E-mail: niko@grigorieff.org
Reason for this suggestion: Expertise in development and application of EM methods for studying structure, function and dynamics of molecular machines.

6) **Andrej Sali**
University of California at San Francisco
UCSF MC 2552, Mission Bay
1700 4th Street,
San Francisco, CA 94158-2330, USA
Email: **sali@salilab.org**
Reason for this suggestion: Expertise in structural bioinformatics methods and applications where the proposed methodology could also be useful.

REVIEWERS TO EXCLUDE (conflict of interest):

1) **Abbas Ourmazd**
Department of Physics
University of Wisconsin
Milwaukee, WI 53211, USA
Email: ourmazd@uwm.edu

2) **Joachim Frank**
HHMI, Columbia University
650 W. 168th Street
New York, NY 10032, USA
Email: jf2192@cumc.columbia.edu

3) **Sjors Scheres**
Medical Research Council
Cambridge Cb2 0qh, United Kingdom
Email: scheres@mrc-lmb.cam.ac.uk

StructMap: Multivariate distance analysis of elastically aligned electron microscopy structures for exploring pathways of conformational changes

C. O. S. Sorzano[1], A. L. Alvarez-Cabrera[1], J. M. Carazo[1], and S. Jonić[2,*]

[1] Biocomputing Unit, Centro Nacional de Biotecnología – CSIC, Campus de Cantoblanco, Darwin 3, 28049 Madrid, Spain.

[2] IMPMC, Sorbonne Universités - CNRS UMR 7590, UPMC Univ Paris 6, MNHN, IRD UMR 206, 75005 Paris, France.

* Corresponding author (Slavica Jonic, IMPMC-UMR 7590, Université Pierre & Marie Curie, Case courrier 115, 4 Place Jussieu, 75005 Paris, France, Phone : +33 1 44 27 72 05, Fax : +33 1 44 27 37 85, E-mail : Slavica.Jonic@impmc.upmc.fr).

Keywords: Structure, dynamics, macromolecular complexes, electron microscopy, conformational changes, dissimilarity, distance, multivariate analysis, pathways

Highlights:

- StructMap maps a given set of EM structures onto a common distance space
- Given structures are visualized as points in a low-dimensional distance space
- Obtained points are analyzed to explore potential conformational-change pathways
- StructMap can be used to initiate full continuous-flexibility image analysis

eTOC blurb:

Sorzano et al. present a new way of looking at the discrete- versus continuous-flexibility problem in molecular electron microscopy (EM), by which a set of distinct EM structures (conformations) is mapped onto a common and quantitative reference frame to reconstitute potential pathways of conformational changes.

Summary: Single-particle electron microscopy has been shown to be very powerful for studying conformational flexibility of macromolecular complexes. To further analyze flexibility and start understanding dynamics of a complex, distinct conformations of this complex are usually computed by analyzing images of its coexisting multiple conformations. The resulting structures are then analyzed to explain flexibility. However, a quantitative analysis of dissimilarities (distances) among structures, placing the entire set of structures into a common space of comparison, is often lacking. Here, we present an approach that provides an overall view of distances among given structures. The approach is based on statistical analysis of distances among elastically aligned structures, and results in visualizing structures as points in a lower-dimensional distance space. The configuration of these points can be analyzed to explore potential pathways of conformational changes. The results of the method are shown with synthetic and experimental structures at different resolutions.

INTRODUCTION

Single-particle analysis (SPA) is routinely used to compute the three-dimensional (3D) structure of a wide range of isolated biological macromolecular complexes (*e.g.*, proteins, ribosomes, viruses), starting from their two-dimensional (2D) transmission electron microscopy (EM) images (Frank, 1996). In this way, EM information, integrated with a large range of other types of data (*e.g.*, from X-ray crystallography, nuclear magnetic resonance, modeling, etc.), often provides very valuable information on how these macromolecular complexes perform their function in the cell.

EM structures are ideally computed from images of complexes having identical conformation and different, uniformly distributed, random orientations. However, quite often, complexes present some degree of flexibility. Methodological extensions of SPA have thus been proposed to analyze flexible complexes (Katsevich et al., 2015, Dashti et al., 2014; Elad et al., 2008; Fu et al., 2007; Jin et al., 2014; Lyumkis et al., 2013a; Penczek et al., 2006; Scheres et al., 2007; Yang et al., 2012), providing very valuable information to understand macromolecular dynamics. A classical approach to analyze macromolecular flexibility is to first classify a set of particle images into distinct classes composed of particles with similar conformations (but random particles orientations), and then to reconstruct a structure for each class (Elad et al., 2008; Fu et al., 2007; Lyumkis et al., 2013a; Penczek et al., 2006; Scheres et al., 2007; Yang et al., 2012). The structures are then analyzed first independently and then with respect to each other, in order to explain the differences between them in terms of conformational flexibility (Fischer et al., 2010; Grob et al., 2006; Lyumkis et al., 2013b; Klinge et al., 2009; Simonetti et al., 2008). However, those classical, class-based approaches, rest on the assumption that flexibility is, indeed, discrete, which is not true for a large range of biological systems characterized by continuous flexibility. Note that should flexibility be a

continuous process and not a discrete one, these class-based approaches will necessarily lead to a resolution loss in each of the class structures, since in reality each structure will be the average of several different, although similar, conformations. It is in this context that new reconstruction approaches, able to explicitly consider continuous flexibility, have been recently proposed (Katsevich et al. 2015, Dashti et al., 2014; Jin et al., 2014). In these approaches, all images from the given set of single particle images are brought into a common and quantitative reference frame in the context of multidimensional analysis of some conformational variables, specific to each of the proposed methods. In this common frame (structure map), images are shown as points and the distance between any two points in the structure map is related to the distance between the corresponding particles (complexes) in terms of their conformational flexibility. Such approaches thus allow analyzing possible pathways of conformational changes by exploring those map regions that are most densely populated.

In this paper, we are presenting a new way of looking at the discrete- versus continuous-flexibility problem. In the proposed approach, a discretization of the flexibility analysis, leading to a distinct set of structures, is followed by a process of elastic alignment of structures and their multidimensional distance analysis, by which all structures are mapped onto a common and quantitative reference frame. The process starts with an elastic volume-to-volume alignment for each pair of EM volumes, using an extension of the elastic volume-to-image alignment method of HEMNMA (Hybrid Electron Microscopy Normal Mode Analysis) (Jin et al., 2014) that was developed in a continuous, normal mode analysis framework (Chacon et al., 2003; Ming et al., 2002; Suhre et al., 2006; Tama et al., 2004; Tama et al., 2002). Then, a matrix of distances among structures (distance matrix) is constructed based on the measured dissimilarities of the elastically aligned volumes. Finally, a

statistical multivariate analysis of the distance matrix is used to map the structures onto a lower-dimensional space (usually, 1D, 2D, or 3D), so that their distances can be visualized in that space. The structures are shown as points in the new space, which allows exploring potential pathways of continuous conformational changes, by analyzing one or several trajectories of three or more connected points. The proposed methodology will be referred to as StructMap, which stands for Structure Mapping.

The results of StructMap are shown with one set of synthetic volumes and two sets of experimental EM density volumes, at different resolutions. The synthetic data were generated using the closed-state structure of the rabbit skeletal muscle type 1 ryanodine receptor (RyR1) complex from Samso et al. (2009). The experimental structures comprise the eukaryotic primosome DNA polymerase Pol α – B subunit complex (Pol α – B) from Klinge et al. (2009) and the *E. coli* 70S ribosome complex from Fischer et al. (2010).

RESULTS

In this section, we first describe the method that we are proposing and, then, we show its performance with synthetic and experimental EM structures.

StructMap design

StructMap comprises the following three steps (**Fig. 1A**): 1) iterative elastic 3D-to-3D alignment of each pair of structures from a given set of EM structures; 2) multivariate analysis of distances among the elastically aligned structures; and 3) analysis and interpretation of the resulting space (graph) of distances among structures (map of structures). We here describe each of these steps.

Iterative elastic 3D-to-3D alignment of EM structures: Given two volumes to be elastically aligned, the alignment is done through rigid-body and elastic displacements of one volume, referred to as "reference volume", until it matches the other volume, referred to as "target volume". The elastic displacement is done by displacing (deforming) the reference volume along a combination of its normal modes. To compute the normal modes, the reference volume is represented with a set of 3D Gaussian functions using a method that controls the volume representation error (Nogales-Cadenas et al., 2013). The 3D Gaussian functions are referred to as "pseudo-atoms", although their positions do not necessarily coincide with the actual atomic positions (Nogales-Cadenas et al., 2013). Also, the volume representation with "pseudo-atoms" is referred to as "pseudo-atomic" structure. A simplified elastic network representation of the potential energy function (Tirion, 1996) is used to compute normal modes of the pseudo-atomic structure. Both pseudo-atoms and normal modes computations are available within the HEMNMA graphical interface (Sorzano et al., 2014) in the open-source software Xmipp (de la Rosa-Trevin et al., 2013; Scheres et al., 2008; Sorzano et al., 2004) and Scipion (de la Rosa-Trevin et al, in preparation).

The iterative alignment method consists in refining amplitudes of the displacement along normal modes (elastic parameters) as well as orientation and position (rigid-body parameters) of the reference volume, by minimizing a measure of dissimilarity between the reference and target volumes. The elastic 3D-to-3D alignment method uses a local-search Powell's UOBYQA optimization that is based on a quadratic-approximation local model of the objective function (Powell, 2002) (**Fig. 1B**). A similar optimization procedure to this one is used in HEMNMA (Jin et al., 2014). Thus, let us here point out the main similarities/differences between the two optimization procedures. As in HEMNMA, in each iteration, the structure is displaced with the current guess of the displacement amplitudes

along normal modes, and the modified pseudo-atomic coordinates are converted into a volume (elastically displaced reference volume). Also, as in HEMNMA, the objective function is evaluated based on the alignment quality. However, the objective function is here evaluated based on the dissimilarity of the best-matching elastically and rigid-body displaced reference volume with respect to the target volume, whereas, in HEMNMA, the displaced reference volume is compared to the target (experimental) images. The rigid-body displacement of the reference volume with respect to the target volume is done with a local fast rotational matching method that maximizes the cross-correlation (*CC*) between the two volumes (Chen et al., 2013) (please note here that the measure of dissimilarity between the two volumes is *S=1-CC*). Finally, the non-deformed reference structure is used to initialize the first iteration (*i.e.*, the initial displacement amplitudes are set to zero), which is also done in HEMNMA.

The alignment method is available in Xmipp. The Xmipp tool (*xmipp_nma_alignment_vol*) aligns a reference volume represented by a pseudo-atomic structure (reference structure) with respect to a set of target volumes. The reference structure is deformed and its orientation and position refined, for each target volume independently, so that the corresponding density volume best matches the target volume.

Multivariate distance analysis: The final (optimal) value of the dissimilarity measure, *S=1-CC*, obtained for each aligned pair of volumes is then used to construct an *N*-by-*N* symmetrical distance matrix *D*, where *N* is the number of given volumes. As the alignment of the *i*-th volume, V_i, with respect to the *j*-th volume, V_j, is done through elastic geometric transformations of V_i until it matches V_j (by displacing the pseudoatomic structure corresponding to V_i along a combination of normal modes, besides a rotation and shift), the

alignment of V_i with respect to V_j will result in dissimilarity, S_{ij}, that, generally, will be different from S_{ji} (the dissimilarity resulting from the alignment of V_j with respect to V_i). Thus, we set the *ij*-th and *ji*-th elements of the distance matrix D (D_{ij} and D_{ji}, respectively) to be the average between S_{ij} and S_{ji} (i.e., $D_{ij}=D_{ji}=(S_{ij}+S_{ji})/2$). We set $D_{ii}=0$ since there is no distance from a volume to itself.

We use a non-metric multidimensional scaling method (Cox and Cox, 2001) to perform multivariate analysis of the distance matrix. The method returns N points in p dimensions, where $p \leq N$ (in normal practice, p=1, p=2 or, at most, p=3), such that the Euclidean distances between the obtained points approximate a monotonic transformation of distances in the distance matrix. The points are plotted in a p-dimensional space that we will refer to as "graph of distances" or "map of structures/structure map".

Analysis and interpretation of the graph of distances (map of structures): The configuration of points in the new, p-dimensional space and their relative distances can be used to identify similar and dissimilar volumes (conformations) and explore possible pathways of conformational changes. In this context, the points may be connected by the user, so as to suggest trajectories that can be explored in terms of potential pathways of conformational changes. The points can be connected in many different ways. Yet, we have found interesting to consider some *ad hoc* alternatives, or "rules", to analyze the data cloud, especially in the absence of *a priori* information about conformational changes of the studied complex. One of such alternatives is to identify the shortest trajectory composed of 3 or more points, two of which are among the most extreme (distant) points (the shortest the trajectory, the lowest is the "energy" or the "effort" required to go from one extreme state to the other). Another possible suggestion is to identify the trajectory composed of points corresponding to those

structures that result from the largest numbers of single particle images (the larger the number of images used to reconstruct a structure, the higher is the probability that the complex adopts this state). These two examples of suggestions can also be combined. Additionally, one could think of analyzing the longest trajectories, so as to explore possible random thermal motions of the complex in solution (Dashti et al., 2014). In this context, one could also consider random Brownian motions around the shortest, minimum-effort paths between extreme states. Though such rules could help to start the analysis, no solid theoretical basis exists to justify a general use of any of them without using additional information (e.g., analysis of overlapped structures).

Experiment 1: Synthetic EM structures

For this experiment, we discretized two distinct synthetic pathways of continuous conformational changes of the same complex. The two continuous conformational pathways were synthesized by displacing a closed-state RyR1 structure (from Samso et al., 2009) along two different normal modes from the same set of normal modes. The two normal modes used for the displacement were selected to produce opening-closing movements of RyR1. The displacement along one of the two modes (mode 8) produces symmetrical conformational changes that correspond to those usually reported in the literature (the structure remains symmetrical after the displacement). The displacement along the other mode (mode 9), in turn, produces asymmetrical conformational changes (the structure becomes asymmetrical after the displacement). To the best of our knowledge, those asymmetrical changes have not been previously reported. Please note here that the two synthesized pathways are completely fictional (they may not exist or coexist in reality).

The goal of the experiment was to analyze the configuration of points in the map of structures obtained using the proposed method and, in particular, to observe how the original conformational pathways look like in this map. In this context, the question was whether the two original conformational pathways look like two distinct trajectories in the structure map, when the points are connected in the order of corresponding structures on the original conformational pathways.

As the synthesized conformational pathways are completely imaginary, it does not make sense to interpret the results of this experiment in terms of potential real conformational pathways of RyR1. Thus, it does not make much sense, either, to try to retrieve the two synthetic "ground-truth" conformational pathways (by connecting points using some of the rules mentioned above), as their retrieval should intimately be linked with their interpretation in terms of potential real conformational changes. Yet, we show here that the retrieval would be possible. Also, we show how we would proceed in an experimental case.

Let us now first fully describe the synthesis of the test data set and, then, present the results of the synthetic data analysis using StructMap.

Synthetic data: The EM structure of the closed-state conformation of RyR1 with code EMD-1606 (resolution: 10.2 Å; volume size: 180^3 voxels; voxel size: $(2.8\ Å)^3$ (Samso et al., 2009)) was converted into pseudo-atoms. This pseudo-atomic structure and its normal modes 8 and 9 were then used to compute 8 additional pseudo-atomic structures. Finally, 500 randomly oriented projections obtained from each of the 9 pseudo-atomic structures were then used to compute the corresponding 3D reconstructions (**Fig. 2A-E**), in all cases considering noise and the effect of the contrast transfer function (CTF) (uniform angular distribution; image

size: 128×128 pixels; pixel size: 3.94 Å × 3.94 Å; defocus: 1 µm; signal-to-noise ratio: 0.2). The CTF and noise were simulated for a 200 kV microscope with a spherical aberration of 2 mm using the method described in (Velazquez-Muriel et al., 2003). Note here that resolution of the reconstructed structures is low (around 20 Å), as a relatively small number (500) of noisy and CTF-affected images was used for the reconstruction. The structure referred to as 1 was reconstructed from projections of the non-displaced pseudo-atomic structure (zero displacement amplitudes along all modes) (**Fig. 2A**). Four of other reconstructed structures, referred to as 2, 3, 4, and 5, were obtained from projections of the pseudo-atomic structure displaced along the mode describing a symmetric opening-closing movement of the complex, using respectively the displacement amplitude of 300, 500, -300, and -500 along mode 8 and zero displacement amplitudes along other modes (note that the corresponding continuous conformational pathway passing through discrete states in the order 3-2-1-4-5 is related to symmetrical opening of the complex, whereas the reverse order 5-4-1-2-3 corresponds to symmetrical closing of the complex) (**Fig. 2B-C**). Four additional synthetic structures, referred to as 6, 7, 8, and 9, were reconstructed from projections of the pseudo-atomic structure displaced along the mode describing an asymmetric opening-closing movement of the complex, using respectively the displacement amplitude of 300, 500, -300, and -500 along mode 9 and zero displacement amplitudes along other modes (note that the corresponding continuous conformational pathway passing through discrete states in the order 7-6-1-8-9 or 9-8-1-6-7 is related to asymmetrical opening or closing of the complex i.e., one side of the complex opens while the other side closes) (**Fig. 2D-E**).

Analysis of synthetic data: A pseudo-atomic structure and its 20 normal modes were computed for each synthetic volume (please see **Experimental Procedures** for an explanation on the choice of the number of modes used for data analysis in **Experiments 1-3**

as well as for other details related to pseudo-atomic structure and normal modes computations). Then, each pseudo-atomic structure was elastically aligned with all EM volumes and the obtained 9-by-9 distance matrix was mapped onto a space of three, two, and one dimensions, respectively (**Figs. 2F-G and 3A-B**). The result of this 3D mapping shows a two-pathway configuration of points 1-9 corresponding to structures 1-9, respectively, when the points are connected in the order of the structures on the original conformational pathways (i.e., 5-4-1-2-3 and 9-8-1-6-7 corresponding to the displacement amplitudes from -500 to 500) (**Fig. 2F**). Also, it shows that differences among neighboring structures on pathway 3-2-1-4-5 are smaller and more similar among each other than differences among neighboring structures on pathway 9-8-1-6-7 (distances among neighboring points on pathway 3-2-1-4-5 are 0.03 in arbitrary units, while they are in the range 0.03-0.06 on pathway 9-8-1-6-7). The different distances among neighboring points on the two pathways mean that the same deformation amplitude has a different impact on the two types of deformations, from the point of view of the mapped dissimilarity measure. The larger distances on pathway 9-8-1-6-7 suggest that the same deformation amplitude has a larger impact on the asymmetrical deformation than on the symmetrical deformation. **Figure 2G** shows that the two sets of points 2-5 and 6-9 are clearly separated, and that structures 7 and 9 are the most distant (different) ones from the rest of structures (*e.g.*, the distance of 0.07 between structures 9 and 2, or between structures 7 and 4).

One can note that trajectory 3-2-1-4-5, related to symmetrical opening of the complex, is the shortest trajectory composed of 3 or more points, two of which are among the most extreme (distant) states (here, maximally closed and maximally open symmetric structures). Also, one can note that trajectory 9-8-1-6-7 (related to asymmetrical opening of the complex) would be identified next, if the same rule was applied. These two trajectories were the ground-truth

synthetic trajectories used in this experiment. In a real data case, other trajectories would also be considered, such as the longest trajectories (e.g., 3-7-2-6-1-8-4-9-5). Indeed, though this experiment was done with synthetic, fictional data, if it was a real experiment, some of states 6-9 could be regarded as "intermediate" states to states 1, 2, or 4, on the way between states 3 and 5 (e.g., by considering random thermal motions of the complex around the minimum-effort (shortest) path between two extreme states (here, states 3 and 5)).

As 2D and 1D mapping could also be useful in practice, especially when analyzing a small number of structures, we mapped the same set of volumes also in these two spaces, which completes our analysis (in practice, the mapping will most often be performed in 1D, 2D, and 3D). The results of 2D mapping are fully consistent with the results of 3D mapping (**Fig. 3A**) and there is a non surprising loss of information about the two-pathway configuration in the case of 1D mapping (**Fig. 3B**). Yet, the 1D mapping still allows to sort out similar and dissimilar structures (*e.g.*, **Fig. 3B** clearly shows that structures 1, 2, and 4 are the most similar ones, whereas structures 7 and 9 are the most dissimilar). 1D mapping may thus still be useful when analyzing a small number of structures (*e.g.*, 3-4 structures), as in the case shown in **Experiment 2**.

Experiment 2: EM structures of Pol α – B

In this experiment, we used three EM structures of Pol α – B published in (Klinge et al., 2009). They correspond to different states of bending of the flexible linker between two lobes of the complex. The EM volumes (volume size: 64 × 64 × 64 voxels; voxel size: 3.8 Å × 3.8 Å × 3.8 Å) have a resolution between 23 Å and 25 Å and are referenced by indexes 1, 2, and 3, as obtained from Klinge et al (2009). A pseudo-atomic structure and its 20 normal modes were first computed for each EM volume (for more details, see **Experimental Procedures**).

Then, each obtained pseudo-atomic structure was elastically aligned with all EM volumes, and the obtained 3-by-3 distance matrix was mapped onto a 1D space (**Fig. 4A**).

The 1D mapping results (**Fig. 4A**) show that structure 1 is almost equally distant from the two other structures (the two distances are 0.1 and 0.11), which could be interpreted as a movement around conformation 1, in the order 3-1-2 or 2-1-3 (**Fig. 4B**). Structures in the order 3-1-2 correspond to the unbending of the complex from conformation 3 to conformation 2 (from left to right in **Figure 4B**). These results are coherent with previously published results (Fig. 6 of Klinge et al (2009)). Moreover, they are based on a quantitative distance analysis, which was not the case in the previous study.

Experiment 3: EM structures of 70S ribosomes

The EM structures corresponding to different pre- and post-translocational states of *E. coli* 70S complex, published by Fischer et al. (2010), were downloaded from EMDB database and, among them, seven structures were selected for the analysis presented here (only structures of the same volume size, 128 × 128 × 128 voxels, the same voxel size, 2.8 Å × 2.8 Å × 2.8 Å, and containing the entire 70S complex, were analyzed). They correspond to the pre-translocational states pre2 to pre5 (EMD-1717 to EMD-1720, respectively) and to the post-translocational states post1 to post3 (EMD-1721 to EMD-1723, respectively), and have different resolutions, between 12 Å and 20 Å (Fischer et al., 2010). A pseudo-atomic structure and its 30 normal modes were first computed for each EM volume (for more details, see **Experimental Procedures**). Then, each obtained pseudo-atomic structure was elastically aligned with all EM volumes and the obtained 7-by-7 distance matrix was mapped onto a 3D space (**Fig. 5A**).

The 3D structure map (**Fig. 5A**) shows that distances between EMDB structures 1717-1720 (often 0.11 and, in general, less than 0.2 arbitrary units apart) are smaller than the distances between these structures and structures 1721-1723 (*e.g.*, the distance between structures 1720 and 1721 is 0.3, the distance of structure 1717 with respect to structures 1721 and 1722 is 0.2 and 0.25, respectively, etc.). This means that structures 1717-1720 correspond to similar conformational states, and that these states are more similar among each other than with respect to states corresponding to other structures, which is also visible when the structures are overlapped. For instance, the orientation of the 30S subunit is more similar between structures 1717-1720 (**Fig. 5B-D**) than is the 30S orientation between these structures and structures 1721-1723 (**Fig. 5G-H**). This is consistent with the findings of Fischer et al. (2010), indicating that structures 1717-1720 correspond to similar, pre-translocational states (pre2 to pre5), whereas structures 1721-1723 correspond to similar, post-translocational states (post1 to post3). Our results complement the results of Fischer et al. (2010) as they present a quantitative analysis of the conformational differences. For instance, the results show that structures 1721 and 1722 are much more similar (the distance is 0.13) than structures 1722 and 1723 (the distance of 0.25) or structures 1721 and 1723 (the distance of 0.25), which is also visible in **Figure 5E-F**.

DISCUSSION

In this paper, we presented StructMap, a methodology that is, to the best of our knowledge, the first one allowing a discrete set of EM structures, obtained by classical discrete (class-based) flexibility analysis, to be represented in a common and quantitative space defined by their mutual distances. The conformational modeling is done by displacing given EM volumes with combinations of normal modes, where the amplitudes of the displacement are computed through elastic alignment among volumes. Elastic alignment allows building a matrix of

distances among volumes, that is then analyzed to map all structures onto a common distance space. The structures are represented in this new space as points, and this space is also referred to as "graph of distances" or "map of structures/structure map". We have shown that such representations could facilitate identifying potential pathways of conformational changes, by exploring one or more trajectories composed of connected points.

Results obtained with synthetic as well as with experimental data show a great potential of the proposed method. Results of the analysis of experimental data of Pol α – B and 70S complexes are fully consistent with previously published results. Moreover, results presented in this paper complement previous results with a quantitative analysis of the configuration of each of the given sets of structures and with a graphical representation of dissimilarities among the structures in the same set.

Potential of the method: Beyond discrete structures

Given a set of EM structures obtained by discretizing flexibility analysis, StructMap can be used to analyze these structures in order to better understand their differences and select a few of them for their use as reference conformations around which the flexibility may be explored more exhaustively, in a fully continuous framework, using techniques such as HEMNMA (Jin et al., 2014; Sorzano et al., 2014) or diffusion map manifold embedding (Dashti et al., 2014). Indeed, such techniques, providing an overall view of the conformational distribution based on a 3D-to-2D alignment of images with a given reference structure (e.g., elastic alignment in the case of HEMNMA, rigid-body alignment in the case of diffusion map manifold embedding), could be used to perform a fine analysis of dynamics around a few conformations selected from the trajectories identified with a help of the proposed methodology. This approach would be less computationally expensive than performing a fine

analysis around every structure from the given set of discrete structures. As StructMap performs an automatic analysis and mapping of structures onto a common distance space, it is especially useful if the given set of discrete structures is large (e.g., containing more than 5 structures; please note that 3-10 structures are typically obtained with discrete methods, though some studies show image analysis resulting in much larger numbers of structures, such as 50 structures (Fischer et al., 2010)). Yet, some human interaction is required for analysis and interpretation of possible conformational trajectories, starting from a discrete set of points in the resulting map of structures, which opens room for further improvements of such combined discrete-continuous approaches.

EXPERIMENTAL PROCEDURES

Computation of pseudo-atomic structures and normal modes

Before the computation of pseudo-atomic structures and normal modes, volumes must be rigid-body aligned, which is a process performed in this work using Xmipp tool *xmipp_volume_align*. The pseudo-atomic structures and their normal modes were computed with HEMNMA graphical interface, by adjusting the following parameters of the protocol described in (Sorzano et al., 2014): 1) a binary volume mask, to avoid representing the structure background noise by Gaussian functions; 2) a target error of the volume representation by Gaussian functions (target volume approximation error), that is a criterion to stop adding Gaussian functions to the pseudo-atomic representation of the volume; 3) the standard deviation of the Gaussian functions; 4) the number of lowest-frequency normal modes to be returned by the interface; 5) the cut-off distance between the pseudo-atoms above which they do not interact in the elastic network model used to compute normal modes; and

6) the modes collectivity threshold determining the set of normal modes to be kept for elastic alignment between volumes (those modes with the collectivity degree below this threshold, as well as the six lowest-frequency modes related to rigid-body movements, are rejected). The binary volume mask was obtained by volume thresholding and several morphological operations. The target value of the volume approximation error was set to a typical value of 5%. The Gaussian-function standard deviation was adjusted to a value between 1 and 2 voxels, which is typically done to better optimize the volume approximation error (note that the interface displays a message to modify the Gaussian standard deviation if the target volume approximation error cannot be achieved). The number of requested normal modes for pseudo-atomic structures is typically between 20 and 50. Here, 20 normal modes were requested for lower-resolution synthetic RyR1 and experimental Pol α – B structures (resolutions lower than 20 Å), whereas 30 normal modes were requested for higher-resolution experimental 70S structures (note here that all 70S structures were treated in the same way, by requesting 30 normal modes, as there was at least one structure of higher resolution (12-14 Å) in the given set of structures). The pseudo-atomic cut-off distance was adjusted to a value between 10 Å and 30 Å, as usually done when using elastic network model. This parameter depends on the size of the complex and the distribution of pseudo-atoms (size of pseudo-atoms as well as their number and distances between them, which are determined by the chosen standard deviation of the Gaussian functions and the achieved volume approximation error). The collectivity threshold was set to a typical value of 0.15, to select only those modes whose collectivities are above this threshold. The number of modes actually used for the elastic alignment may thus be different for different pseudo-atomic structures.

ACKNOWLEDGMENTS

The work was partially funded by the CNRS (France) and the CSIC (Spain) [Projet International de Coopération Scientifique - PICS 2011]; the French National Research Agency ANR [ANR-11-BSV8-010-04]; the European Social Fund and the Ministerio de Educación y Ciencia ["Ramón y Cajal" fellowship to COSS]; the Spanish Ministry of Economy and Competitiveness [AIC–A–2011–0638 and BIO2013-44647-R]; and the Comunidad de Madrid [CAM S2010/BMD-2305]. We thank GENCI-CINES/IDRIS (France) for HPC resources [x2013072174, x2014072174, x2015072174], and L. Pellegrini, S. Klinge (Cambridge University, UK) and O. Llorca (CIB-CSIC, Spain) for generously providing the Pol α – B EM structures.

Conflict of interest statement. None declared.

REFERENCES

Chacon, P., Tama, F., and Wriggers, W. (2003). Mega-Dalton biomolecular motion captured from electron microscopy reconstructions. J Mol Biol *326*, 485-492.

Chen, Y., Pfeffer, S., Hrabe, T., Schuller, J. M., and Forster, F. (2013). Fast and accurate reference-free alignment of subtomograms. J Struct Biol *182*, 235-245.

Cox, T. F., and Cox, M. A. A. (2001). Multidimensional Scaling, 2nd edn (Boca Raton, FL: Chapman & Hall/CRC).

Dashti, A., Schwander, P., Langlois, R., Fung, R., Li, W., Hosseinizadeh, A., Liao, H. Y., Pallesen, J., Sharma, G., Stupina, V. A.*, et al.* (2014). Trajectories of the ribosome as a Brownian nanomachine. Proc Natl Acad Sci U S A *111*, 17492-17497.

de la Rosa-Trevin, J. M., Oton, J., Marabini, R., Zaldivar, A., Vargas, J., Carazo, J. M., and Sorzano, C. O. (2013). Xmipp 3.0: an improved software suite for image processing in electron microscopy. J Struct Biol *184*, 321-328.

Elad, N., Clare, D. K., Saibil, H. R., and Orlova, E. V. (2008). Detection and separation of heterogeneity in molecular complexes by statistical analysis of their two-dimensional projections. J Struct Biol *162*, 108-120.

Fischer, N., Konevega, A. L., Wintermeyer, W., Rodnina, M. V., and Stark, H. (2010). Ribosome dynamics and tRNA movement by time-resolved electron cryomicroscopy. Nature *466*, 329-333.

Frank, J. (1996). Three-Dimensional Electron Microscopy of Macromolecular Assemblies (San Diego: Academic Press).

Fu, J., Gao, H., and Frank, J. (2007). Unsupervised classification of single particles by cluster tracking in multi-dimensional space. J Struct Biol *157*, 226-239.

Grob, P., Cruse, M. J., Inouye, C., Peris, M., Penczek, P. A., Tjian, R., and Nogales, E. (2006). Cryo-Electron Microscopy Studies of Human TFIID: Conformational Breathing in the Integration of Gene Regulatory Cues. Structure *14*, 511-520.

Jin, Q., Sorzano, C. O., de la Rosa-Trevin, J. M., Bilbao-Castro, J. R., Nunez-Ramirez, R., Llorca, O., Tama, F., and Jonic, S. (2014). Iterative elastic 3D-to-2D alignment method using normal modes for studying structural dynamics of large macromolecular complexes. Structure *22*, 496-506.

Katsevich, E., Katsevich, A., and Singer, A. (2015). Covariance Matrix Estimation for the Cryo-EM Heterogeneity Problem. SIAM J Imaging Sci *8*, 126-185.

Klinge, S., Nunez-Ramirez, R., Llorca, O., and Pellegrini, L. (2009). 3D architecture of DNA Pol alpha reveals the functional core of multi-subunit replicative polymerases. Embo J *28*, 1978-1987.

Lyumkis, D., Brilot, A. F., Theobald, D. L., and Grigorieff, N. (2013a). Likelihood-based classification of cryo-EM images using FREALIGN. J Struct Biol *183*, 377-388.

Lyumkis, D., Doamekpor, S. K., Bengtson, M. H., Lee, J. W., Toro, T. B., Petroski, M. D., Lima, C. D., Potter, C. S., Carragher, B., and Joazeiro, C. A. (2013b). Single-particle EM reveals extensive conformational variability of the Ltn1 E3 ligase. Proc Natl Acad Sci U S A *110*, 1702-1707.

Ming, D., Kong, Y., Lambert, M. A., Huang, Z., and Ma, J. (2002). How to describe protein motion without amino acid sequence and atomic coordinates. Proceedings of the National Academy of Sciences *99*, 8620-8625.

Nogales-Cadenas, R., Jonic, S., Tama, F., Arteni, A. A., Tabas-Madrid, D., Vazquez, M., Pascual-Montano, A., and Sorzano, C. O. (2013). 3DEM Loupe: Analysis of macromolecular dynamics using structures from electron microscopy. Nucleic Acids Res *41*, W363-367.

Penczek, P. A., Frank, J., and Spahn, C. M. (2006). A method of focused classification, based on the bootstrap 3D variance analysis, and its application to EF-G-dependent translocation. J Struct Biol *154*, 184-194.

Powell, M. J. D. (2002). UOBYQA: unconstrained optimization by quadratic approximation. Mathematical Programming *92*, 555-582.

Samso, M., Feng, W., Pessah, I. N., and Allen, P. D. (2009). Coordinated movement of cytoplasmic and transmembrane domains of RyR1 upon gating. PLoS Biol *7*, e85.

Scheres, S. H., Gao, H., Valle, M., Herman, G. T., Eggermont, P. P., Frank, J., and Carazo, J. M. (2007). Disentangling conformational states of macromolecules in 3D-EM through likelihood optimization. Nat Methods *4*, 27-29.

Scheres, S. H., Nunez-Ramirez, R., Sorzano, C. O., Carazo, J. M., and Marabini, R. (2008). Image processing for electron microscopy single-particle analysis using XMIPP. Nat Protoc *3*, 977-990.

Simonetti, A., Marzi, S., Myasnikov, A. G., Fabbretti, A., Yusupov, M., Gualerzi, C. O., and Klaholz, B. P. (2008). Structure of the 30S translation initiation complex. Nature *455*, 416-420.

Sorzano, C. O., de la Rosa-Trevin, J. M., Tama, F., and Jonic, S. (2014). Hybrid Electron Microscopy Normal Mode Analysis graphical interface and protocol. J Struct Biol *188*, 134-141.

Sorzano, C. O., Marabini, R., Velazquez-Muriel, J., Bilbao-Castro, J. R., Scheres, S. H., Carazo, J. M., and Pascual-Montano, A. (2004). XMIPP: a new generation of an open-source image processing package for electron microscopy. J Struct Biol *148*, 194-204.

Suhre, K., Navaza, J., and Sanejouand, Y. H. (2006). NORMA: a tool for flexible fitting of high-resolution protein structures into low-resolution electron-microscopy-derived density maps. Acta Crystallogr D Biol Crystallogr *62*, 1098-1100.

Tama, F., Miyashita, O., and Brooks, C. L., 3rd (2004). Flexible multi-scale fitting of atomic structures into low-resolution electron density maps with elastic network normal mode analysis. J Mol Biol *337*, 985-999.

Tama, F., Wriggers, W., and Brooks, C. L., 3rd (2002). Exploring global distortions of biological macromolecules and assemblies from low-resolution structural information and elastic network theory. J Mol Biol *321*, 297-305.

Tirion, M. M. (1996). Large Amplitude Elastic Motions in Proteins from a Single-Parameter, Atomic Analysis. Phys Rev Lett *77*, 1905-1908.

Velazquez-Muriel, J. A., Sorzano, C. O., Fernandez, J. J., and Carazo, J. M. (2003). A method for estimating the CTF in electron microscopy based on ARMA models and parameter adjustment. Ultramicroscopy *96*, 17-35.

Yang, Z., Fang, J., Chittuluru, J., Asturias, F. J., and Penczek, P. A. (2012). Iterative stable alignment and clustering of 2D transmission electron microscope images. Structure *20*, 237-247.

FIGURE LEGENDS

Figure 1: Flowchart of the proposed method StructMap (A) and the iterative elastic 3D-to-3D alignment step (between any two volumes) used in the method (B). The measure of dissimilarity between two finally aligned volumes is taken as the distance between the two volumes. The distances among all pairs of volumes are used to construct a distance matrix that is then analyzed with a multivariate analysis method, so as to map all volumes onto a common distance space, in which each volume is represented as a point. Points may then be

connected, suggesting trajectories that can be explored in terms of potential pathways of conformational changes.

Figure 2: Synthetic RyR1 data experiment. (A) Synthetic volume 1. (B-C) Synthetic volumes representing symmetrical structural changes (volume 1 (grey) overlapped with volumes 2 (yellow) and 3 (cyan) (B); volume 1 overlapped with volumes 4 (violet) and 5 (magenta) (C)). (D-E) Synthetic volumes representing asymmetrical structural changes (volume 1 (grey) overlapped with volumes 6 (rose) and 7 (green) (D); volume 1 overlapped with volumes 8 (brown) and 9 (blue) (E)). (F-G) Two views of the mapping of structures onto a 3D distance space. In A-E, all volumes are viewed from the same viewing direction. In F-G, the structures are marked with their indexes and circles. The dotted lines are used to show distances between the structures that are discussed in the text. The length of each line segment (the distance between two conformations) is shown above the segment in arbitrary units. See also **Figure 3**.

Figure 3: Mapping of synthetic RyR1 structures onto a distance space of a lower dimension than 3. (A) 2D mapping. (B) 1D mapping. The structures are marked with their indexes and circles. See also **Figure 2**.

Figure 4: Mapping of three EM structures of DNA polymerase Pol α - B complex onto a 1D distance space and analysis of these structures based on their distances in the new space. (A) Mapping of structures onto a 1D distance space. (B) Ordered rigid-body aligned structures (3-1-2) according to their distances in the distance space shown in A. In A, the structures are marked with circles and the length of each dotted line segment (the distance between two structures) is shown above the segment (in arbitrary units).

Figure 5: Mapping of seven 70S ribosome structures (EMD-1717 to EMD-1723) onto a 3D distance space and analysis of these structures based on their distances in the new space. (A) Mapping of structures onto a 3D distance space. (B-H) Overlap of rigid-body aligned structures, with the same view for all overlapped structures (volumes 1718, 1719, and 1720 *vs* volume 1717 (gray) in B, C, and D, respectively; volumes 1722 and 1723 *vs* volume 1721 (magenta) in E and F, respectively; and volumes 1717 and 1720 *vs* volume 1721 (magenta) in G and H, respectively). In A, the structures are marked with circles and the corresponding EMDB entry codes. Straight lines are used to connect neighboring conformations, according to the original proposal in Fischer et al. (2010), and the length of each line segment (the distance between two conformations) is shown above the segment (in arbitrary units). The dotted lines are used to show additional distances that are discussed in the text.

Figure 1
Click here to download Figure: Fig1_Jonic.jpg

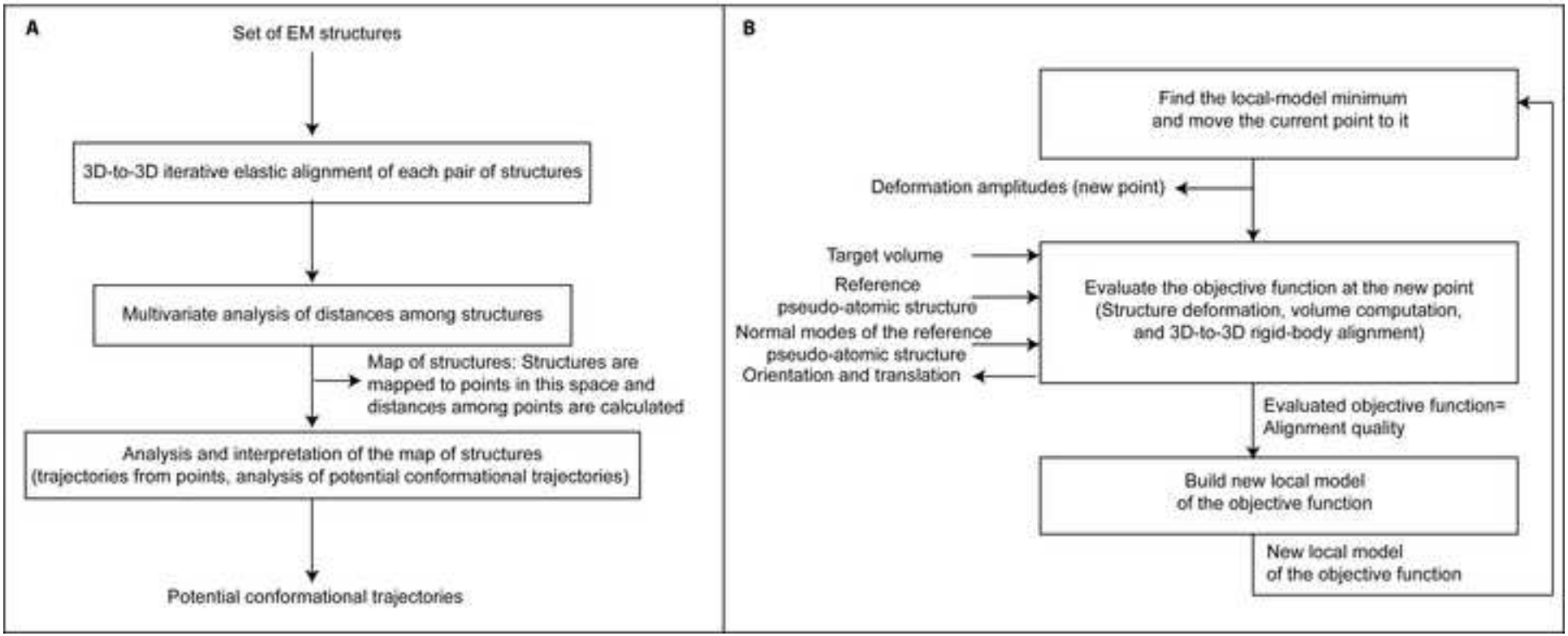


Figure 2
Click here to download Figure: Fig2_Jonic_RGB.jpg

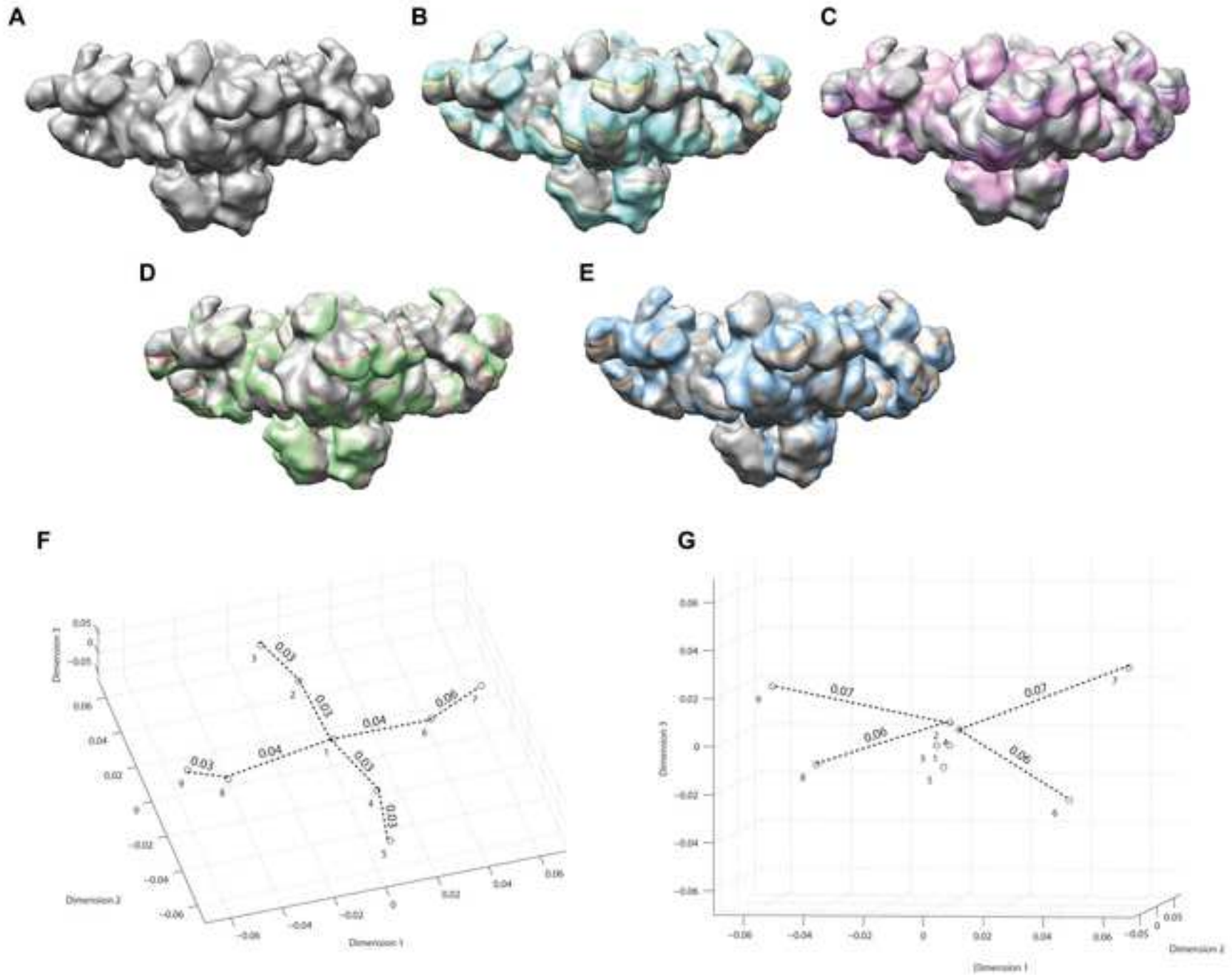
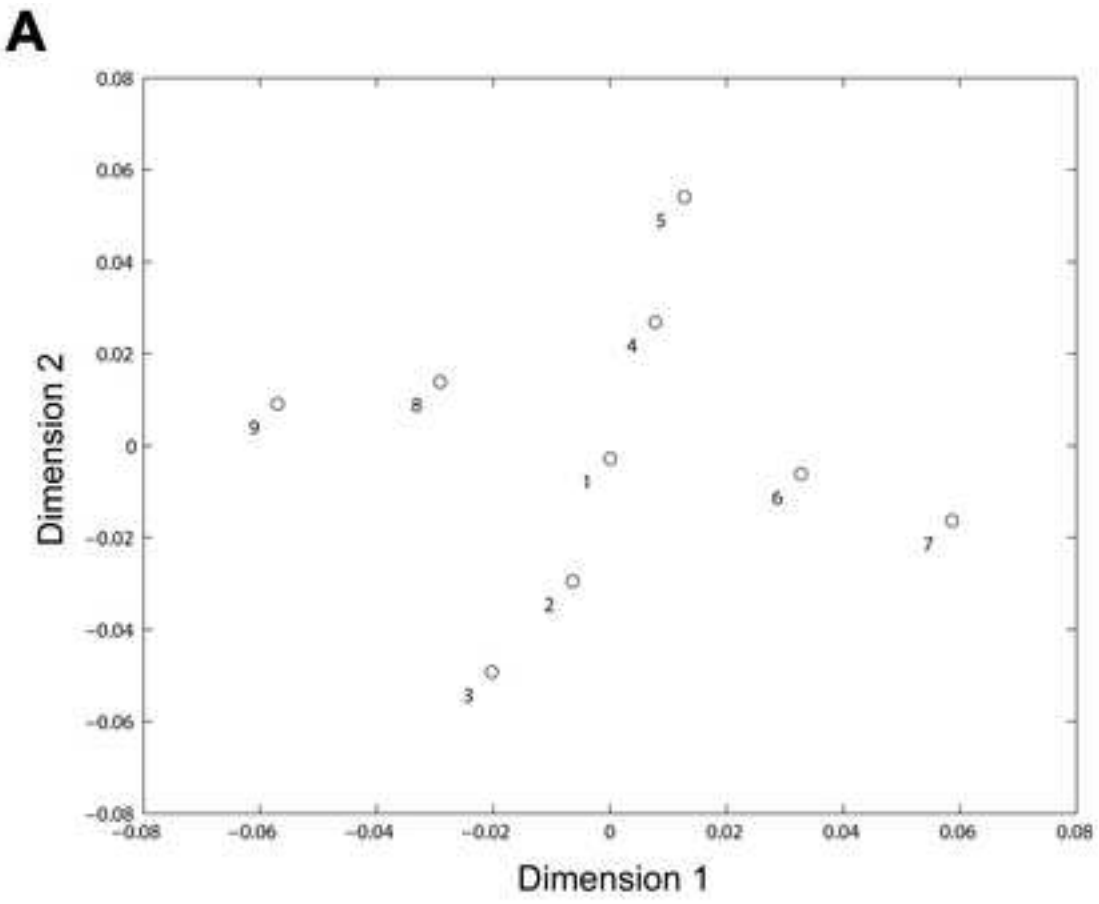


Figure 3
Click here to download Figure: Fig3_Jonic.jpg



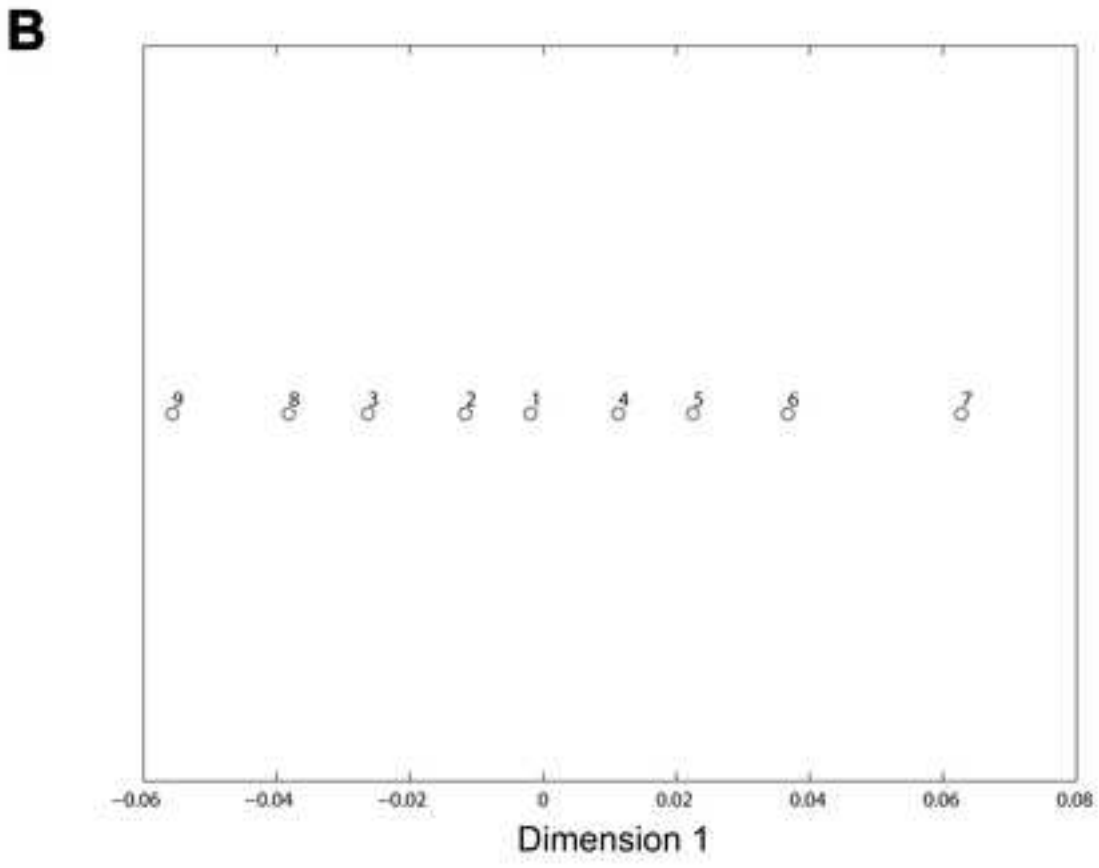


Figure 4
Click here to download Figure: Fig4_Jonic_RGB.jpg

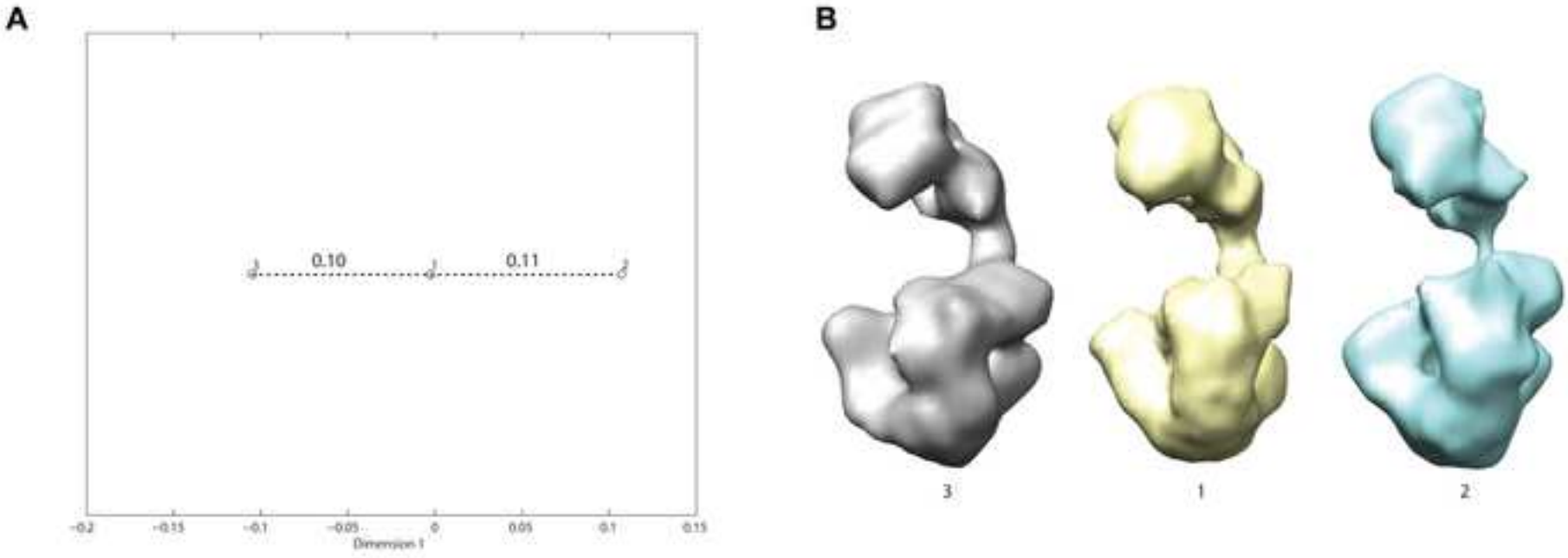


Figure 5
Click here to download Figure: Fig5_Jonic_RGB.jpg